# Bayesian Higher Order Hidden Markov Models

Abhra Sarkar
Department of Statistics and Data Sciences,
University of Texas at Austin,
2317 Speedway D9800, Austin, TX 78712-1823, USA
abhra.sarkar@utexas.edu
and
David B. Dunson
Department of Statistical Science,
Duke University,
Box 90251, Durham NC 27708-0251
dunson@duke.edu

# **Abstract**

We consider the problem of flexible modeling of higher order hidden Markov models when the number of latent states and the nature of the serial dependence, including the true order, are unknown. We propose flexible Bayesian methods based on tensor factorization techniques that can characterize any transition probability with a specified maximal order, allowing automated selection of the important lags and capturing higher order interactions among the lags. Theoretical results provide insights into identifiability of the emission distributions and consistency of the posterior. We design Markov chain Monte Carlo algorithms for posterior computation. In simulation experiments, the method vastly outperforms competitors not just in higher order settings, but, remarkably, also in first order cases. Practical utility is illustrated using real world applications.

**Some Key Words**: Bayesian nonparametrics, Conditional tensor factorization, Higher order hidden Markov models, Sequential data, Time series.

Short/Running Title: Higher Order Hidden Markov Models

Correpsonding Author: Abhra Sarkar (abhra.sarkar@utexas.edu)

# 1 Introduction

Hidden Markov models (HMMs) have been tremendously successful in statistical analyses of sequentially generated data (Frühwirth-Schnatter, 2006; McDonald and Zuchhini, 1997; Cappé et al., 2005) in diverse application areas like proteomics (Bae et al., 2005; Lennox et al., 2010), genomics (Guha et al., 2008; Yau et al., 2011; Titsias et al., 2016), animal movement (Langrock et al., 2015; Quick et al., 2017), speech recognition (Rabiner, 1989; Fox et al., 2011), and economics and finance (Hamilton, 1990; Albert and Chib, 1993).

The basic HMM consists of two processes: a *hidden* process  $\{c_t\}$ , which evolves according to a first order Markov chain with discrete state space, and a potentially multivariate *observed* process  $\{\mathbf{y}_t\}$  observed sequentially over a set of discrete time points t = 1, 2, ..., T. Specifically, an HMM makes the following set of conditional independence assumptions to model the hidden and the observed processes

$$p(c_t \mid \mathbf{c}_{1:(t-1)}) = p(c_t \mid c_{t-1}), \tag{1}$$

$$p(\mathbf{y}_t \mid \mathbf{y}_{1:(t-1)}, \mathbf{c}_{1:(t-1)}) = p(\mathbf{y}_t \mid c_t).$$
 (2)

The distributions  $p(c_t \mid c_{t-1})$  and  $p(\mathbf{y}_t \mid c_t)$  are often referred to as the transition distribution and the emission distribution, respectively.

A challenging problem of the HMM framework is the determination of the cardinality of the state space. This is often unknown in practice and is determined using model selection approaches (Sclove, 1983; Leroux and Puterman, 1992; Wang and Puterman, 1999) or reversible jump type model space exploration techniques (Robert et al., 2000). Teh et al. (2006) developed a Bayesian nonparametric approach to HMMs based on the hierarchical Dirichlet process (HDP) that defines a prior distribution on transition matrices over a countably infinite number of states. The number of latent states for any given dataset can be inferred from its posterior, allowing for uncertainty in the analysis and also the possibility that additional states may be required when more data points become available, precluding the necessity to decide a priori the size of the state space.

One serious limitation of the HDP-HMM in particular and the basic HMM framework in general is the restrictive assumption of first order Markovian dynamics of the latent sequence  $\{c_t\}$ . The focus of this article is on higher order HMMs (HOHMMs) that allow  $\{c_t\}$  to depend on its more distant past. An HOHMM of maximal order q thus makes the following set of conditional independence assumptions

$$p(c_t \mid \mathbf{c}_{1:(t-1)}) = p(c_t \mid \mathbf{c}_{(t-q):(t-1)}),$$
 (3)

$$p(\mathbf{y}_t \mid \mathbf{y}_{1:(t-1)}, \mathbf{c}_{1:(t-1)}) = p(\mathbf{y}_t \mid c_t). \tag{4}$$

We distinguish between an HOHMM of maximal order q and an HOHMM of full order q. An HOHMM is said to be of maximal order q if conditional on the values of  $c_{t-1}, \ldots, c_{t-q}$ , the distribution of  $c_t$  is independent of its more distant past, but the lags actually important in determining the distribution of  $c_t$  may be an arbitrary subset of  $\{c_{t-1}, \ldots, c_{t-q}\}$ . In contrast, if the distribution of  $c_t$  actually varies with the values at all the previous q times points, we call the HOHMM to be of full order q. The case q = 0 corresponds to serial independence of the observation sequence  $\{\mathbf{y}_t\}$ . Also, we say that an HOHMM of maximal order q has true maximal order q, if the set of important predictors of  $c_t$  includes  $c_{t-q}$ .

While the HOHMM framework relaxes the restrictive first order assumption of the basic HMM, it also brings in a daunting dimensionality challenge. Consider, for instance, an HOHMM with C states and maximal order q. The transition distributions are now indexed by the  $C^q$  different possible values of the lags  $\mathbf{c}_{(t-q):(t-1)}$  (rather than just  $c_{t-1}$ ), and involve a total number of  $(C-1)C^q$  parameters, which increases exponentially with the order and becomes too large to be estimated efficiently with datasets of the sizes typically encountered in practice. The issue is further complicated by the fact that we do not directly observe the values of the latent sequence  $\{c_t\}$  but only their noisy manifestations  $\{\mathbf{y}_t\}$ .

Any HOHMM can be reformulated as a first order HMM by moving either in blocks of q time steps or, more conventionally, in single time steps but with a special q-tuple initial latent state (McDonald and Zuchhini, 1997; Cappé  $et\ al.$ , 2005). While convenient for theoretical treatment of HOHMMs, for modeling purposes such formulations are not very useful since they require working with large  $C^q \times C^q$  dimensional transition probability matrices with  $C^q(C^q-C)$  structural zeros. Associated computational machineries also quickly become practically ineffective even for moderately small values of C and q. Parsimonious characterization of the transition dynamics in higher order settings is thus extremely important. It is also important to obtain an interpretable structure, with unnecessary lags eliminated.

These daunting challenges to higher order generalizations have forced researchers to focus on first order HMMs. Thede and Harper (1999) used a second order HMM for parts of speech tagging, estimating the transition probabilities by weighted mixtures of empirical proportions of subsequences of maximal length three. Seifert  $et\ al.\ (2012)$  developed an HOHMM with known finite state space and Normal emission densities for modeling array comparative genomic hybridization (aCGH) data. Transition dynamics of maximal order q were modeled using state context trees of maximal depth q that divide the set of all possible state combination histories into disjoint sets of equivalent state contexts. Tree-based strategies employ strict top-down search for important lags and hence are not suitable for scenarios when distant lags may be more important than recent ones (Jääskinen  $et\ al.\ 2014$ ; Sarkar and Dunson, 2016).

In this article, we develop a novel Bayesian nonparametric approach to HOHMMs

that can parsimoniously characterize the transition dynamics of any HOHMM with a specified maximal order, allows flexibility in modeling the emission distributions, admits generalizations to countably infinite state spaces, precluding the necessity to predetermine the number of states, and allows automated selection of the important lags, determining the true order and nature of the serial dependence, removing the necessity to decide *a priori* the exact order of the transition dynamics.

We begin by structuring the transition probabilities  $p(c_t \mid \mathbf{c}_{(t-q):(t-1)})$  as a high dimensional conditional probability tensor. Adapting the conditional tensor factorization approach of Yang and Dunson (2016) to the HOHMM setting, we parameterize the probabilities  $p(c_t \mid \mathbf{c}_{(t-q):(t-1)})$  as mixtures of 'core' probability kernels with mixture weights depending on the state combinations of the lags. Such a parameterization explicitly identifies the set of important lags and implicitly captures complex higher order interactions among the important lags, borrowing strength across the states of the HOHMM by sharing the core kernels in a 'soft' probabilistic manner. The elimination of the redundant lags and the implicit modeling of the interactions among the important ones can lead to a significant two fold reduction in the effective number parameters required to flexibly characterize the transition dynamics of the HOHMM. We assign sparsity inducing priors that favor such lower dimensional representations of the transition probability tensor.

We assign a hierarchical Dirichlet prior on the core probability kernels, encouraging the model to shrink further towards lower dimensional structures by borrowing strength across these components as well. This also facilitates a generalization to countably infinite state space HOHMMs that allow uncertainty in the number of states. The HDP-HMM of Teh *et al.* (2006) corresponds to a special case when the kernel sharing feature is turned off and the order is restricted to one.

We develop a two-stage Markov chain Monte Carlo (MCMC) algorithm for learning the parameters of the model. The first stage selects the important lags implementing a coarser 'hard' sharing approximation using a stochastic search variable selection (SSVS) approach (George and McCulloch, 1997). The second stage keeps the set of important lags fixed and implements the finer soft kernel sharing feature, building on existing computational machineries for the HDP-HMM.

HOSVD-type factorizations have previously been employed in Sarkar and Dunson (2016) to model the transition dynamics of observable state sequences in a higher order Markov chain framework. The framework of HOHMM, however, brings in significant additional challenges. Unlike an observable Markov process, the states  $c_t$  are now latent, only their noisy manifestations  $\mathbf{y}_t$  are available. The size of the state space is often unknown and has to be inferred from these noisy data points. These issues make infinite state space models particularly relevant in the HOHMM context. The emission distributions  $p(\mathbf{y}_t \mid c_t)$  have to be additionally modeled which brings in identifiability issues and significant computational challenges.

The rest of the article is organized as follows. Section 2 details the proposed tensor factorization based HOHMMs and their properties. Section 3 describes Markov chain Monte Carlo (MCMC) algorithms for drawing samples from the posterior. Section 4 presents the results of simulation experiments comparing our method with existing approaches. Section 5 presents some real world applications. Section 6 contains concluding remarks.

# 2 Higher Order Hidden Markov Model

# 2.1 Modeling the Transition Probabilities

We build on the idea of higher order singular value decomposition (HOSVD) tensor factorization to develop a nonparametric approach for modeling the transition dynamics of a finite memory HOHMM. HOSVD (Tucker, 1966; De Lathauwer *et al.*, 2000) factorizes a  $C_1 \times \cdots \times C_p$  dimensional p-way tensor  $\mathbf{M} = \{m_{x_1,\dots,x_p}\}$  as

$$m_{x_1,\dots,x_p} = \sum_{h_1=1}^{k_1} \dots \sum_{h_p=1}^{k_p} g_{h_1,\dots,h_p} \prod_{j=1}^p u_{h_j x_j}^{(j)},$$

where the core tensor  $\mathbf{G} = \{g_{h_1,\dots,h_p}\}$  captures the interactions between different components and  $\mathbf{U}^{(j)} = \{u_{h_j x_j}^{(j)}\}$  are component specific weights.

In our HOHMM setting, the hidden sequence  $\{c_t\}$  with state space  $\{1, \ldots, C\}$  has finite memory of true maximal order q. Given  $c_{t-q}, \ldots, c_{t-1}$ , the distribution of  $c_t$  is independent of all latent states prior to t-q. The variables that are important in predicting  $c_t$  comprise a subset of  $\{c_{t-q}, \ldots, c_{t-1}\}$ , possibly proper but including  $c_{t-q}$ .

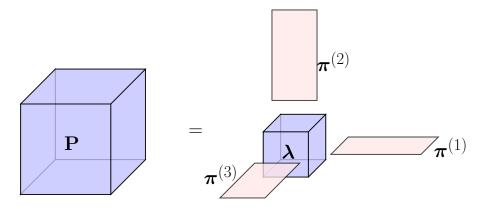


Figure 1: Pictorial representation of the factorization of a transition probability tensor **P** characterizing a Markov chain of maximal order 3 with core tensor  $\lambda$  and mode matrices  $\pi^{(j)}$ , j = 1, 2, 3.

We structure the transition probabilities  $p(c_t \mid c_{t-q}, \dots, c_{t-1})$  as a  $C \times C \times \dots \times C$  dimensional (q+1)-way tensor and consider the following HOSVD-type factorization

$$p(c_t \mid c_{t-j}, j = 1, \dots, q) = \sum_{h_1=1}^{k_1} \cdots \sum_{h_q=1}^{k_q} \lambda_{h_1, \dots, h_q}(c_t) \prod_{j=1}^q \pi_{h_j}^{(j)}(c_{t-j}).$$
 (5)

See Figure 1. Here  $1 \leq k_j \leq C$  for all j and the parameters  $\lambda_{h_1,\dots,h_q}(c_t)$  and  $\pi_{h_j}^{(j)}(c_{t-j})$  are all non-negative and satisfy the constraints (a)  $\sum_{c_t=1}^C \lambda_{h_1,\dots,h_q}(c_t) = 1$ , for each combination  $(h_1,\dots,h_q)$ , and (b)  $\sum_{h_j=1}^{k_j} \pi_{h_j}^{(j)}(c_{t-j}) = 1$ , for each pair  $(j,c_{t-j})$ . If follows as a direct consequence of Theorem 1 in Yang and Dunson (2016), a general result on conditional probability tensors, that any transition probability tensor can be represented as (5) with the parameters satisfying the constraints (a) and (b).

Introducing latent allocation variables  $z_{j,t}$  for  $j=1,\ldots,q$  and  $t=q+1,\ldots,T$ , the latent variables  $\{c_t\}$  are conditionally independent and the factorization can be equivalently represented through the following hierarchical formulation

$$(c_t \mid z_{j,t} = h_j, j = 1, \dots, q) \sim \text{Mult}(\{1, \dots, C\}, \lambda_{h_1, \dots, h_q}(1), \dots, \lambda_{h_1, \dots, h_q}(C)), (6)$$

$$(z_{j,t} \mid c_{t-j}) \sim \text{Mult}(\{1, \dots, k_j\}, \pi_1^{(j)}(c_{t-j}), \dots, \pi_{k_j}^{(j)}(c_{t-j})). (7)$$

See Figure 2. Equation (7) reveals the soft sharing property of the model that enables it to borrow strength across the different states of  $c_{t-j}$  by allowing the  $z_{j,t}$ 's associated with a particular state of  $c_{t-j}$  to be allocated to different latent populations, which are shared across all C states of  $c_{t-j}$ . In contrast, a hard sharing model would allocate each  $z_{t,j}$  to a single latent population. Equation (6) shows how such soft assignment enables the model to capture complex interactions among the lags in an implicit and parsimonious manner by allowing the latent populations indexed by  $(h_1, \ldots, h_q)$  to be shared among the various state combinations of the lags.

When  $k_j = 1$ ,  $\pi_1^{(j)}(c_{t-j}) = 1$  and  $P(c_t \mid c_{t-q}, \dots, c_{t-1})$  does not vary with  $c_{t-j}$ . The variable  $k_j$  thus determines the inclusion of the  $j^{th}$  lag  $c_{t-j}$  in the model. The variable  $k_j$  also determines the number of latent classes for the  $j^{th}$  lag  $c_{t-j}$ . The number of parameters in such a factorization is given by  $(C-1)\prod_{j=1}^q k_j + C\sum_{j=1}^q (k_j-1)$ , which will be much smaller than the number of parameters  $(C-1)C^q$  required to specify a full Markov model of the same maximal order, if  $\prod_{j=1}^q k_j \ll C^q$ .

As a first proposal, we may assign independent Dirichlet priors on  $\lambda_{h_1,...,h_q}$  as

$$\lambda_{h_1,\dots,h_q} \sim \text{Dir}(\alpha,\dots,\alpha)$$
, independently for each  $(h_1,\dots,h_q)$ . (8)

The estimation of  $\prod_{j=1}^{q} k_j$  independent parameters may still be a daunting task in practical applications. Even in first order HMMs, single layer independent priors

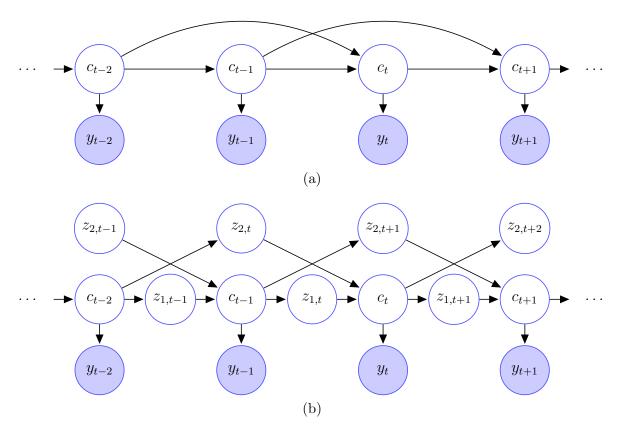


Figure 2: Graphical model depicting the dependence structure of a second order hidden Markov model (a) without and (b) with second level latent variables. Shaded and unshaded nodes represent observed and latent variables, respectively.

on the transition probability vectors have far inferior performance to hierarchical priors that allow information to be better shared between different state-dependent transition distributions. See, for example, Figure 10 in Fox *et al.* (2011).

Towards a more parsimonious representation of the transition probability tensor, we assign a conditionally independent hierarchical Dirichlet prior on  $\lambda_{h_1,...,h_q} = \{\lambda_{h_1,...,h_q}(1),...,\lambda_{h_1,...,h_q}(C)\}$ . Specifically, we let

$$\lambda_{h_1,\dots,h_q} \sim \text{Dir}\{\alpha\lambda_0(1),\dots,\alpha\lambda_0(C)\}, \text{ independently for each } (h_1,\dots,h_q),$$
 (9)

$$\lambda_0 = \{\lambda_0(1), \dots, \lambda_0(C)\} \sim \operatorname{Dir}(\alpha_0/C, \dots, \alpha_0/C). \tag{10}$$

The 'kernels'  $\lambda_{h_1,\dots,h_q}$  are associated with the mixture weights in a hierarchical DP. The dimension of  $\pi_{k_j}^{(j)}(c_{t-j}) = \{\pi_1^{(j)}(c_{t-j}),\dots,\pi_{k_j}^{(j)}(c_{t-j})\}$ , unlike the  $\lambda_{h_1,\dots,h_q}$ 's, varies only linearly with  $k_j$ . We assign independent priors on the  $\pi_{k_j}^{(j)}(c_{t-j})$ 's as

$$\boldsymbol{\pi}_{k_j}^{(j)}(c_{t-j}) = \{ \pi_1^{(j)}(c_{t-j}), \dots, \pi_{k_j}^{(j)}(c_{t-j}) \} \sim \operatorname{Dir}(\gamma_j, \dots, \gamma_j).$$
 (11)

While the dimension of the core tensor varies with  $k_j$ 's, all  $\lambda_{h_1,\dots,h_q}$  share the same support  $\{1,\dots,C\}$ . This allows us to avoid conditioning on the  $k_j$ 's while specifying the hierarchical prior on  $\lambda_{h_1,\dots,h_q}$ . The probability vectors  $\boldsymbol{\pi}_{k_j}^{(j)}(c_{t-j})$ , on the other hand, are supported on  $\{1,\dots,k_j\}$  for each pair  $(j,c_{t-j})$ . Therefore, unlike  $\lambda_{h_1,\dots,h_q}$ , conditioning on  $k_j$ , which we have kept implicit in (11), can not be avoided.

Finally, we assign the following independent priors on  $k_i$ 's

$$p_{0,j}(k) \propto \exp(-\varphi jk), \quad j = 1, \dots, q, \quad k = k_{j,\min}, \dots, C,$$
 (12)

where  $\varphi > 0$ ,  $k_{j,\min} = 1$  for  $j = 1, \ldots, q-1$  and  $k_{q,\min} = 2$ . The prior  $p_{0,j}$  assigns increasing probabilities to smaller values of  $k_j$  as the lag j becomes more distant, reflecting the natural belief that increasing lags have diminishing influence on the distribution of  $c_t$ . The larger the value of  $\varphi$ , the faster is the decay of  $p_{0,j}(k)$  with increase in j and k, favoring sparser lower order models. The restriction  $k_q > 1$  implies that the  $q^{th}$  lag is important so that the true maximal order is q.

### 2.2 Modeling the Emission Distributions

The generic form of the emission distribution that we consider in this article is

$$p(\mathbf{y}_t \mid c_t, \boldsymbol{\beta}, \boldsymbol{\psi}) = f(\mathbf{y}_t \mid \boldsymbol{\beta}_{c_t}, \boldsymbol{\psi}).$$

Here  $\boldsymbol{\beta} = \{\boldsymbol{\beta}_c : c = 1, ..., C\}$  denotes parameters indexed by the latent process  $\{c_t\}$ , whereas  $\boldsymbol{\psi}$  collects global parameters that do not evolve with time but remain constant and may sometimes be kept implicit.

In the first order HMM literature, parametric choices for the emission distribution are common. Leroux (1992) provided sufficient conditions for identifiability in such models. There has been some recent interest in flexible models for the emission distributions (Yau et al., 2011; Langrock et al., 2015) that relax restrictive parametric assumptions, which can be shown to satisfy sufficient conditions for identifiability (Gassiat et al., 2015; Alexandrovich et al., 2016). The following lemma establishes such guarantees in higher order settings.

**Lemma 1.** Let P be the transition probability tensor and  $\mathbf{f}$  be the emission distributions of an HOHMM with known state space  $\mathcal{C}$  and known true maximal order q. Let the first order representation of the underlying Markov chain be ergodic and stationary with transition probability matrix  $\widetilde{P}$  and stationary and initial distribution  $\{\pi(c_1,\ldots,c_q):c_j\in\mathcal{C},j=1,\ldots,q\}$ . Let  $\widetilde{P}$  be of full rank and the emission distributions  $\mathbf{f}=\{f_c:c\in\mathcal{C}\}$  be all distinct. Then P and  $\mathbf{f}$  are nonparametrically identifiable from the distribution of  $T=(2C^q+1)q$  consecutive observations  $\mathbf{y}_{1:T}$  up to label swapping of the states.

The proof, deferred to section S.1 in the Supplementary Materials, utilizes a similar result on first order HMMs from Alexandrovich *et al.* (2016), noting that by moving in blocks of size q as

$$(c_1, \dots, c_q) \rightarrow (c_{q+1}, \dots, c_{2q}) \rightarrow (c_{2q+1}, \dots, c_{3q}) \rightarrow \cdots$$

$$\downarrow \qquad \qquad \downarrow \qquad \qquad \downarrow$$

$$(\mathbf{y}_1, \dots, \mathbf{y}_q) \qquad (\mathbf{y}_{q+1}, \dots, \mathbf{y}_{2q}) \qquad (\mathbf{y}_{2q+1}, \dots, \mathbf{y}_{3q}) \qquad \cdots$$

an HOHMM of maximal order q with state space  $\mathcal{C} = \{1, \ldots, C\}$  and transition probability tensor P can be represented as a first order HMM with expanded state space  $\mathcal{C}^q$ , stationary and initial distribution  $\pi(c_1, \ldots, c_q)$ , emission distributions  $\{f_{c_1}f_{c_2}\cdots f_{c_q}: c_j \in \mathcal{C}, j = 1, \ldots, q\}$ , and q-step transition probability matrix  $\widetilde{P}^q$ , where the single-step transition probabilities are given by

$$\widetilde{P}\{(j_{t-q}, \dots, j_{t-1}), (i_{t-q+1}, \dots, i_t)\} 
= p\{(c_{t-q+1} = i_{t-q+1}, \dots, c_t = i_t) \mid (c_{t-q} = j_{t-q}, \dots, c_{t-1} = j_{t-1})\} 
= \begin{cases}
P(c_t = i_t \mid c_{t-q} = j_{t-q}, \dots, c_{t-1} = j_{t-1}), & \text{if } i_{t-\ell} = j_{t-\ell} \text{ for } \ell = 1, \dots, (q-1), \\
0, & \text{otherwise.} 
\end{cases}$$

Lemma 1 assumes nonsingularity of  $\widetilde{P}$ . This does not limit its applicability to HOHMMs of full orders but also accommodates lag gaps. In this case, the transition probability matrix  $\widetilde{P}$  will have multiple rows sharing the same nonzero elements but they will appear in different columns so that  $\widetilde{P}$  could still have full rank. Consider, for example, a binary Markov chain of maximal order 2 with a lag gap at t-1 so that  $P(c_t \mid c_{t-2}, c_{t-1}) = P(c_t \mid c_{t-2})$  and  $\widetilde{P}$  is given by

$$\widetilde{P} = \begin{array}{cccc} (1,1) & (1,2) & (2,1) & (2,2) \\ (1,1) & \begin{bmatrix} P(1 \mid 1) & P(2 \mid 1) & 0 & 0 \\ 0 & 0 & P(1 \mid 1) & P(2 \mid 1) \\ (2,1) & [P(1 \mid 2) & P(2 \mid 2) & 0 & 0 \\ 0 & 0 & P(1 \mid 2) & P(2 \mid 2) \end{bmatrix}.$$

The implication of the restriction  $k_q > 1$  in (12) in ensuring nonsingularity of  $\widetilde{P}$  is now clear.

A result on identifiability of HOHMMs in parametric settings can be derived along the lines of Leroux (1992) where only ergodicity of  $\tilde{P}$  suffices. For such choices, an unrestricted independent prior on the  $k_j$ 's would suffice and the restriction  $k_q > 1$  may be dropped. Treating q to be an upper bound on the maximal order, the proposed model can then select the important lags itself, including zeroth order cases which can be viewed as HOHMMs with  $P(c_t \mid c_{t-q}, \ldots, c_{t-1}) = \pi(c_t)$ . In applications of HOHMMs, however, some form of serial dependency would generally be expected and we do not pursue the zeroth order cases any further. Practical strategies that allow the assumption of known true maximal order to be relaxed are discussed in Section 2.6.

In this article, we consider the following families of emission distributions - (a) Normal, (b) Poisson and (c) translated mixture of Normals. For Gaussian emission distributions  $f(y \mid c_t = c) = \text{Normal}(y \mid \mu_c, \sigma_c^2)$ . We assign conjugate Normal $(\mu_0, \sigma_0^2) \times \text{Inv-Ga}(a_0, b_0)$  priors on  $(\mu_c, \sigma_c^2)$ . For Poisson emission distributions  $f(y \mid c_t = c) = \text{Poi}(y \mid \mu_c)$ , we assign conjugate  $\text{Ga}(a_0, b_0)$  prior on  $\mu_c$ , with the hyper-parameters chosen such that  $E(\mu_c) = a_0 b_0 = \overline{y}$ ,  $\text{var}(\mu_c) = a_0 b_0^2 = 2 \text{var}(y)$ . Finally, translated mixture Normal emission distributions are constructed as

$$f(y \mid c_t = c) = \sum_{s=1}^{S} \pi_s \text{Normal}(y \mid \mu_c + \eta_s, \sigma_s^2), \text{ subject to } \sum_{s=1}^{S} \pi_s \eta_s = 0.$$

Introducing additional latent variables  $s_t \in \{1, ..., S\}$  for each t, the model can be rewritten hierarchically as

$$f(y \mid c_t = c, s_t = s) = \text{Normal}(y \mid \mu_c + \eta_s, \sigma_s^2), \quad p(s_t = s) = \pi_s.$$

The states  $s_t$ 's model local departures from the state specific means  $\mu_c$ 's but are globally shared across all states. The moment restriction ensures that the marginal mean of each latent state c is still  $\mu_c$ . The model is similar to that in Yau et al. (2011), where they did not have any moment restriction on the globally shared components  $\mu_s$  but assumed one local mean  $\mu_c$  to be exactly known to identify the state specific means. We assign the priors  $\mu_c \sim \text{Normal}(\mu_0, \sigma_0^2)$ ,  $\pi_{\eta} = (\pi_1, \dots, \pi_S)^{\text{T}} \sim \text{Dir}(\alpha_{\eta}/S, \dots, \alpha_{\eta}/S)$ ,  $\eta_s \sim \text{Normal}(\mu_{\eta,0}, \sigma_{\eta,0}^2)$ , and  $\sigma_s^2 \sim \text{Inv-Ga}(a_0, b_0)$ .

#### 2.3 Likelihood Factorization

Collecting all potential predictors of  $c_t$  in  $\mathbf{w}_t = (w_{1,t}, \dots, w_{q,t})^{\mathrm{T}}$  with  $w_{j,t} = c_{t-j}$  for  $j = 1, \dots, q$  and  $t = t^*, \dots, T$ , where  $t^* = (q+1)$ , the joint distribution of  $\mathbf{y} = \{\mathbf{y}_t : t = 1, \dots, T\}$ ,  $\mathbf{c} = \{c_t : t = t^*, \dots, T\}$  and  $\mathbf{z} = \{z_{j,t} : t = t^*, \dots, T, j = 1, \dots, q\}$  admits the following factorization

$$p(\mathbf{y}, \mathbf{c}, \mathbf{z} \mid \boldsymbol{\lambda}_{\mathbf{k}}, \boldsymbol{\pi}_{\mathbf{k}}, \mathbf{k}) = \prod_{t=t^{\star}}^{T} \left\{ p(c_{t} \mid \boldsymbol{\lambda}_{\mathbf{z}_{t}}) \prod_{j=1}^{q} p(z_{j,t} \mid w_{j,t}, \boldsymbol{\pi}_{k_{j}}^{(j)}, k_{j}) \right\} \prod_{t=1}^{T} f(\mathbf{y}_{t} \mid \boldsymbol{\beta}_{c_{t}}, \boldsymbol{\psi})$$

$$= \prod_{t=1}^{T} f(\mathbf{y}_{t} \mid \boldsymbol{\beta}_{c_{t}}, \boldsymbol{\psi}) \prod_{t=t^{\star}}^{T} \{ p(c_{t} \mid \boldsymbol{\lambda}_{\mathbf{z}_{t}}) p(\mathbf{z}_{t} \mid \mathbf{w}_{t}, \boldsymbol{\pi}_{\mathbf{k}}, \mathbf{k}) \}$$

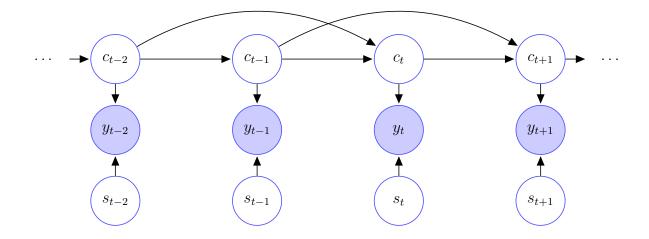


Figure 3: Graphical model depicting the dependence structure of a second order hidden Markov model with translated mixtures as emission distributions.

$$= p(\mathbf{y} \mid \mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\psi}) p(\mathbf{c} \mid \mathbf{z}, \boldsymbol{\lambda}_{\mathbf{k}}, \mathbf{k}) \prod_{j=1}^{q} p(\mathbf{z}_{j} \mid \mathbf{w}_{j}, \boldsymbol{\pi}_{k_{j}}^{(j)}, k_{j})$$

$$= p(\mathbf{y} \mid \mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\psi}) p(\mathbf{c} \mid \mathbf{z}, \boldsymbol{\lambda}_{\mathbf{k}}, \mathbf{k}) p(\mathbf{z} \mid \mathbf{w}, \boldsymbol{\pi}_{\mathbf{k}}, \mathbf{k}),$$
(13)

Here  $\mathbf{k} = \{k_j : j = 1, \dots, q\}$ ,  $\lambda_{\mathbf{k}} = \{\lambda_{h_1,\dots,h_q}(c) : c = 1,\dots,C,h_j = 1,\dots,k_j, j = 1,\dots,q\}$ ,  $\pi_{k_j}^{(j)}(w_j) = \{\pi_{h_j}^{(j)}(w_j) : h_j = 1,\dots,k_j\}$ ,  $\pi_{k_j}^{(j)} = \{\pi_{k_j}^{(j)}(w_j) : w_j = 1,\dots,C\}$ ,  $\pi_{\mathbf{k}} = \{\pi_{k_j}^{(j)} : j = 1,\dots,q\}$ . Also,  $\mathbf{z}_t = \{z_{j,t} : j = 1,\dots,q\}$  for all  $t = t^*,\dots,T$ ,  $\mathbf{z}_j = \{z_{j,t} : t = t^*,\dots,T\}$  for  $j = 1,\dots,q$  and  $\mathbf{w}_j = \{w_{j,t} : t = t^*,\dots,T\}$ . The subscripts  $\mathbf{k}$  and  $k_j$  signify that the dimensions of the associated parameters depend on them. In what follows, the subscript  $\mathbf{k}$  may sometimes be dropped from  $\lambda_{\mathbf{k}}$  to highlight that, unlike  $\pi_{\mathbf{k}}$ , the support of the core probability vectors comprising  $\lambda_{\mathbf{k}}$  does not depend on  $\mathbf{k}$ . The conditional independence relationships encoded in the factorization are used in deriving MCMC algorithms to draw samples from the posterior in Section 3.

# 2.4 Posterior Consistency

Consistency of the posterior of Bayesian first order HMMs under the frequentist assumption that there is a true fixed parameter has been studied in Vernet (2015a,b); Gassiat and Rousseau (2014). Asymptotic treatment of the posterior is facilitated under the assumptions of stationarity and uniform ergodicity of the underlying Markov chain, and some additional mild assumptions on the emission distributions. Specifically, it is assumed that the true transition probabilities as well as their priors are all bounded below by some positive number  $\underline{p}$  (Vernet, 2015a,b). Under similar assumptions on  $p(c_t \mid \mathbf{c}_{(t-q):(t-1)})$ , such results can be extended to HOHMMs.

In the following, we write  $\pi_{c_1,\dots,c_q} = \pi(c_1,\dots,c_q)$ ,  $p_{c_{q+1}|c_1,\dots,c_q} = p(c_{q+1} \mid c_1,\dots,c_q)$  and  $f(\cdot \mid c) = f_c(\cdot)$ . With some abuse of notation, we let  $\boldsymbol{\theta} = (\mathbf{p},\mathbf{f}) \in \mathcal{P} \times \mathcal{F} = \Theta$  collect the parameters of the model, where  $\mathcal{P}$  denotes the space of transition probabilities  $\mathbf{p}$  and  $\mathcal{F}$  denotes the space of emission distributions  $\mathbf{f} = (f_1,\dots,f_C)^T$  with respect to some reference measure  $\lambda$  on the observation space  $\mathcal{Y} \subset \mathbb{R}^d$  for some positive integer d. Let  $\Pi = \Pi_P \times \Pi_F$ , where  $\Pi_P$  and  $\Pi_F$  denote the priors on the transition probabilities and the emission distribution, respectively. Let  $\Pi(\cdot \mid \mathbf{y}_{1:T})$  denote the corresponding posterior based on an observed sequence  $\mathbf{y}_{1:T}$  of length T. Let

$$f_{1:\ell}^{\star}(\mathbf{y}_{1},\ldots,\mathbf{y}_{\ell}\mid\boldsymbol{\theta}) = \sum_{c_{1}=1}^{C}\cdots\sum_{c_{\ell}=1}^{C}\pi_{c_{1},\ldots,c_{q}}\prod_{t=q+1}^{\ell}p_{c_{t}\mid c_{(t-q)},\ldots,c_{(t-1)}}\prod_{t=1}^{\ell}f_{c_{t}}(\mathbf{y}_{t}),$$

denote the  $\ell^{th}$  order marginal distribution of a stationary HOHMM. Let

$$D_{\ell}(\boldsymbol{\theta}_{0},\boldsymbol{\theta}) = \int |f_{1:\ell}^{\star}(\mathbf{y}_{1},\ldots,\mathbf{y}_{\ell} \mid \boldsymbol{\theta}_{0}) - f_{1:\ell}^{\star}(\mathbf{y}_{1},\ldots,\mathbf{y}_{\ell} \mid \boldsymbol{\theta})| \lambda(d\mathbf{y}_{1}) \cdots \lambda(d\mathbf{y}_{\ell}).$$

For  $\mathbf{f}, \mathbf{f}_0 \in \mathcal{F}$ , let  $d(\mathbf{f}, \mathbf{f}_0) = \max_c \int |f_c(\mathbf{y}) - f_{c,0}(\mathbf{y})| \lambda(d\mathbf{y})$  and  $\mathcal{N}(\delta, \mathcal{F}, d)$  is the  $\delta$ -covering number of the set  $\mathcal{F}$  with respect to the distance metric d. Let  $\mathcal{P}(\underline{p})$  denote the space of transition probability tensors supported on the compact set  $\underline{p} \leq p(c_t \mid c_{(t-q),\dots,(t-1)}) \leq 1 - (C-1)\underline{p}$  for some  $0 < \underline{p} < 1/C$ . In the proposed tensor factorized formulation (5), such a restriction can be imposed by assigning truncated Dirichlet priors on  $\lambda_{h_1,\dots,h_q}$ .

**Assumption 1.** a. The true data generating process is a stationary HOHMM of maximal order q with parameter  $\boldsymbol{\theta}_0 = (\mathbf{p}_0, \mathbf{f}_0) \in \Theta(\underline{p}) = \mathcal{P}(\underline{p}) \times \mathcal{F}$ .

b. For all  $\epsilon > 0$ , there exists  $\Theta_{\epsilon} = \mathcal{P}_{\epsilon} \times \mathcal{F}_{\epsilon} \subset \Theta(p)$  such that for all  $(\mathbf{p}, \mathbf{f}) \in \Theta_{\epsilon}$ 

b.1. 
$$\Pi(\Theta_{\epsilon}) = \Pi_{P}(\mathcal{P}_{\epsilon})\Pi_{F}(\mathcal{F}_{\epsilon}) > 0$$
,

b.2. 
$$\max_{\mathbf{c}_{(t-q):t}} \left| p_{c_t|c_{(t-q)},\dots,c_{(t-1)}} - p_{c_t|c_{(t-q)},\dots,c_{(t-1)},0} \right| < \epsilon,$$

b.3. 
$$\max_{c} \int f_{c,0}(\mathbf{y}) \max_{k} \log \frac{f_{k,0}(\mathbf{y})}{f_{k}(\mathbf{y})} \lambda(d\mathbf{y}) < \epsilon$$
,

b.4. 
$$\sum_{c} f_c(\mathbf{y}) > 0$$
 whenever  $\sum_{c} f_{c,0}(\mathbf{y}) > 0$ ,

b.5. 
$$\sup_{\{\mathbf{y}: \sum_{c} f_{c,0}(\mathbf{y}) > 0\}} \max_{k} f_{k}(\mathbf{y}) < \infty$$
,

b.6. 
$$\sum_{c} \int f_{c,0}(\mathbf{y}) |log\{\sum_{k} f_{k}(\mathbf{y})\}| \lambda(d\mathbf{y}) < \infty$$
,

c. For all T and all  $\epsilon > 0$ , there exists  $\mathcal{F}_T \subset \mathcal{F}$  such that

c.1. 
$$\Pi_F(\mathcal{F}_T^c) \leq \exp(-T\beta_1)$$
 for some  $\beta_1 > 0$ ,

c.2. 
$$\sum_{T=1}^{\infty} \mathcal{N}\left\{\frac{\epsilon}{36\ell}, \mathcal{F}_T, d(\cdot, \cdot)\right\} \exp\left(-\frac{T\epsilon^2 C^2 \underline{p}^2}{32\ell}\right) < \infty.$$

**Theorem 1.** Under Assumptions 1, for  $\ell \geq q$ , for any  $\epsilon > 0$ ,

$$P_{\boldsymbol{\theta}_0} \left[ \lim_{T \to \infty} \Pi\{\boldsymbol{\theta} : D_{\ell}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) < \epsilon \mid \mathbf{y}_{1:T}\} = 1 \right] = 1.$$

The proof of the theorem and some additional convergence results are discussed in Section S.2 in the Supplementary Materials.

#### 2.5 Prediction

For a  $q^{th}$  order HMM with state space C, transition probabilities  $p(c_t \mid c_{t-q}, \dots, c_{t-1})$  and emission distributions  $\{f(y \mid c) : c \in C\}$ , the r-step ahead density is

$$f_{pred,T+r}(y) = \sum_{c_{T+r}} \sum_{c_{T+r-1}} \cdots \sum_{c_{T+1}} f(y \mid c_{T+r}) p(c_{T+r} \mid \mathbf{c}_{(T+r-q):(T+r-1)}) \dots p(c_{T+1} \mid \mathbf{c}_{(T+1-q):T}).$$

With stationary distribution  $\boldsymbol{\pi} = \{\pi(c_{t-q+1}, \dots, c_t) : c_j \in \mathcal{C}, j = t-q+1, \dots, t\}$ , for  $r \to \infty$  we then have

$$\sum_{c_{T+r-q}} \cdots \sum_{c_{T+1}} p(c_{T+r} \mid \mathbf{c}_{(T+r-q):(T+r-1)}) \dots p(c_{T+1} \mid \mathbf{c}_{(T+1-q):T}) \to \pi(\mathbf{c}_{(T+r-q+1):(T+r)}).$$

The marginal probabilities of occurrences of individual states  $i \in \mathcal{C}$ , denoted with slight abuse of notation also by  $\pi(i)$ , may be obtained from  $\pi$  by fixing the last (or any other) element in  $\pi$  at i and then summing across the values of the remaining entries. That is,  $\pi(i) = \sum_{c_{t-q+1},\dots,c_{t-1}} \pi(c_{t-q+1},\dots,c_{t-1},i)$ . Likewise, for any  $(i,j) \in \mathcal{C}^2$ , we have  $\pi(i,j) = \sum_{c_{t-q+1},\dots,c_{t-2}} \pi(c_{t-q+1},\dots,c_{t-2},i,j)$ . This implies, as  $r \to \infty$ 

$$\sum_{c_{T+r-1}} \cdots \sum_{c_{T+1}} p(c_{T+r} \mid \mathbf{c}_{(T+r-q):(T+r-1)}) \dots p(c_{T+1} \mid \mathbf{c}_{(T+1-q):T})$$

$$\to \sum_{c_{T+r-1}} \cdots \sum_{c_{T+r-q+1}} \pi(\mathbf{c}_{(T+r-q+1):(T+r)}) = \pi(c_{T+r}).$$

Hence, we have

$$f_{pred,T+r}(y) \to \sum_{c} \pi(c) f(y \mid c).$$

Next, consider a first order HMM, characterized by the transition probabilities  $\{P(j \mid i) = \pi(i,j)/\pi(j) : i,j \in \mathcal{C}\}$ , stationary distribution  $\{\pi(i) : i \in \mathcal{C}\}$  and emission distributions  $\{f(y \mid c) : c \in \mathcal{C}\}$ . The r-step ahead density then approaches the same

limit as  $r \to \infty$ . That is, we have

$$f_{pred,T+r}(y) = \sum_{c_{T+r}} \sum_{c_{T+r-1}} \cdots \sum_{c_{T+1}} f(y \mid c_{T+r}) P(c_{T+r} \mid c_{T+r-1}) \dots P(c_{T+1} \mid c_{T}) \to \sum_{c} \pi(c) f(y \mid c).$$

As will be seen in Section 4, significant gains in efficiency in estimating several steps ahead predictive densities can be achieved through modeling higher order dynamics when such lags are truly present. As the number of steps ahead is increased, the performances of higher and comparable first order HMMs in estimating the predictive densities, will, however, tend to be similar. In both cases, as r increases, the error in estimating  $f_{pred,T+r}(y)$  will also tend to stabilize.

### 2.6 Unknown Maximal Order and Unknown State Space

In Sections 2.1-2.5, we assumed the maximal order and the size of the state space to be known. In practical applications, one or both of these quantities are often unknown. In this section, we devise practical strategies to relax these assumptions.

We first relax the assumption of known maximal order, letting q to be a known upper bound on the maximal order, and using the following prior.

$$p(k_1, \dots, k_q) \propto 1 \left\{ \sum_{j=1}^q k_j > q \right\} \prod_{j=1}^q p_{0,j}(k),$$
 (14)

$$p_{0,j}(k) \propto \exp(-\varphi jk), \quad j = 1, \dots, q, \quad k = 1, \dots, C.$$
 (15)

The restriction  $k_q > 1$  is now replaced by  $\sum_{j=1}^q k_j > q$  which ensures that at least one lag is important and the transition matrix corresponding to the true order has full rank. The proposed methodology automatically selects the important lags.

To relax the assumption of known state space, we look toward Bayesian nonparametric models that can accommodate countably infinitely many states in the prior and allow the number of states required to model the data to be sampled and inferred from the posterior. Such models also accommodate the possibility that additional latent states may be required to allow the model to grow in complexity as more data points become available. To this end, the finite state space model for the latent sequence  $\{c_t\}$  proposed in Section 2.1 is extended to

$$(c_t \mid z_{j,t} = h_j, j = 1, \dots, q) \sim \lambda_{h_1,\dots,h_q},$$
 (16)

$$\lambda_{h_1,\dots,h_q} \sim \mathrm{DP}(\alpha,\lambda_0), \qquad \lambda_0 \sim \mathrm{SB}(\alpha_0),$$
 (17)

$$(z_{j,t} \mid c_{t-j}) \sim \boldsymbol{\pi}_{k_j}^{(j)}(c_{t-j}), \qquad \boldsymbol{\pi}_{k_j}^{(j)}(c_{t-j}) \sim \text{Dir}(\gamma_j, \dots, \gamma_j),$$
 (18)

$$p(k_1, \dots, k_q) \propto 1 \left\{ \sum_{j=1}^q k_j > q \right\} \prod_{j=1}^q p_{0,j}(k), \qquad p_{0,j}(k_j) \propto \exp(-\varphi j k_j).$$
 (19)

Here  $h_j = 1, ..., k_j; k_j = 1, ..., \infty; j = 1, ..., q$ ;  $\mathrm{DP}(\alpha, \lambda_0)$  denotes a Dirichlet process prior (Ferguson, 1973) with concentration parameter  $\alpha$  and base probability measure  $\lambda_0$ ; and  $\lambda_0 \sim \mathrm{SB}(\alpha_0)$  denotes the stick-breaking construction (Sethuraman, 1994) of  $\lambda_0 = \{\lambda_0(1), \lambda_0(2), ...\}$  as

$$\lambda_0(\ell) = v_\ell \prod_{m=1}^{\ell-1} (1 - v_m), \ v_\ell \sim \text{Beta}(1, \alpha_0), \ \ell = 1, 2, \dots$$

Equation (17) defines an HDP prior on the probability distributions  $\lambda$ . In the special case of a first order HMM, with q = 1, the model reduces to

$$(c_t \mid z_t = h) \sim \lambda_h, \tag{20}$$

$$\lambda_h \sim \mathrm{DP}(\alpha, \lambda_0) \text{ for } h = 1, \dots, k, \qquad \lambda_0 \sim \mathrm{SB}(\alpha_0)$$
 (21)

$$(z_t \mid c_{t-1}) \sim \boldsymbol{\pi}_k(c_{t-1}), \qquad \boldsymbol{\pi}_k(c_{t-1}) \sim \operatorname{Dir}(\gamma_j, \dots, \gamma_j),$$
 (22)

$$p_0(k) \propto \exp(-\varphi k), \quad k = 2, \dots, \infty.$$
 (23)

The HDP-HMM of Teh et al. (2006) is obtained as a further special case if we let  $k = \infty$ , and  $\pi_h(c) = 1$  if h = c and 0 otherwise for all  $h = 1, ..., \infty$ . The proposed model thus generalizes the HDP-HMM in at least two directions. First, it models higher order transition dynamics that can also accommodate lag gaps. Second, even in the special first order setting, the soft allocation feature of the model, as opposed to the hard clustering in HDP-HMMs, enables better representation of the dynamics, resulting in improved estimation and prediction performance. See Section 4.

For moderate to large values of C, the finite dimensional prior  $\operatorname{Dir}(\alpha/C,\ldots,\alpha/C)$  provides a weak limit approximation to the infinite dimensional Dirichlet process prior (Ishwaran and Zarepour, 2002) in the sense that if  $G_C = \sum_{\ell=1}^C \lambda_\ell \delta_{\theta_k}$  with  $\lambda \sim \operatorname{Dir}(\alpha/C,\ldots,\alpha/C)$  and  $\theta_k \sim G_0$ , then, for any measurable function g integrable with respect to  $G_0$ ,  $\int g(\theta)dG_C(\theta) \stackrel{d}{\to} \int g(\theta)dG(\theta)$  as  $C \to \infty$ , where  $G \sim \operatorname{DP}(\alpha,G_0)$ . The finite state-space HOHMM model proposed in Section 2.1 thus provides an excellent practical basis for approximate inference on integrable functionals of the infinite dimensional model (16)-(19). In effect, as in the case of maximal order, having a known finite upper bound on the state space size suffices.

# 3 Posterior Computation

In this article, inference about the proposed HOHMM is based on samples drawn from the posterior using MCMC algorithms. In our proposed HOHMM, the values of  $k_j$ 's, being crucial in controlling the model size and acting as lag selection indicators, are of inferential importance. Varying values of  $k_j$ 's, however, result in varying dimensional models, posing significant computational challenges. Dynamic message passing algorithms, such as the forward-backward sampler, are popular strategies for inference in first order HMMs. See Rabiner (1989) for a review and Scott (2002) for Bayesian MCMC based adaptations. More conventional strategies can be found in Frühwirth-Schnatter (2001). Such algorithms cannot, however, be straightforwardly adapted to the HOHMM setting. It is not clear how message passing strategies can be adapted to include inferences about the  $k_j$ 's. Even if the  $k_j$ 's are known, when higher order lags are present, computing forward or backward messages would require summing across all possible past histories comprising all possible combinations of values of important lags for each state at each time stamp at each iteration. This involves a prohibitively large number of operations.

We address these challenges by designing a two-stage MCMC algorithm. In the first stage, we sample the  $k_j$ 's from the posterior of a coarser version of the proposed model. This coarser version is itself fully capable of modeling any transition probability tensor with the  $k_j$ 's still being interpretable as lag selection indicators. In the second stage, we sample from the posterior keeping the  $k_j$ 's fixed. In what follows, the notation  $\zeta$  is used to collect all parameters and data points that are not explicitly mentioned.

The first stage proceeds as follows. The mixture probabilities are now denoted by  $\widetilde{\boldsymbol{\pi}}_{k_j}$ , for reasons to become obvious shortly. Given the current values of  $\mathbf{k}$  and  $\mathbf{c}$ , we partition the levels of  $w_{j,t} = c_{t-j}$  into  $k_j$  clusters  $\{\mathcal{C}_{j,r} : r = 1, \ldots, k_j\}$  with each cluster  $\mathcal{C}_{j,r}$  assumed to correspond to its own latent class  $h_j = r$ . The cluster mappings are then defined as  $\widetilde{\pi}_{h_j}^{(j)}(c_{t-j}) = 1$  for  $h_j = r$  and  $\widetilde{\pi}_{h_j}^{(j)}(c_{t-j}) = 0$  otherwise. This imposes restrictions on soft allocation of the  $z_{t,j}$ 's, forcing the coarser hard allocation instead. With  $\lambda_{h_1,\ldots,h_q} \sim \text{Dir}\{\alpha\lambda_0(1),\ldots,\alpha\lambda_0(C)\}$  marginalized out, conditional on the cluster configurations  $\mathcal{C} = \{\mathcal{C}_{j,r} : j = 1,\ldots,q,r = 1,\ldots,k_j\}$ , we then have

$$p(\mathbf{c} \mid \mathcal{C}, \boldsymbol{\zeta}) = \prod_{(h_1, \dots, h_q)} \frac{\beta \{\alpha \lambda_0(1) + n_{h_1, \dots, h_q}(1), \dots, \alpha \lambda_0(C) + n_{h_1, \dots, h_q}(C)\}}{\beta \{\alpha \lambda_0(1), \dots, \alpha \lambda_0(C)\}}, (24)$$

where  $n_{h_1,\dots,h_q}(c) = \sum_{t=t^*}^T 1\{c_t = c, w_{1,t} \in \mathcal{C}_{1,h_1},\dots,w_{q,t} \in \mathcal{C}_{m,h_q}\}$ . We then use an SSVS approach (George and McCulloch, 1997) based on the approximated marginal likelihood (24) to sample the  $k_j$ 's from their posterior. Conditional on  $\mathbf{k}$  and the current cluster mappings, we then update  $\mathbf{c}$  and  $\mathbf{z}$  using a Metropolis-Hastings step.

#### Algorithm 1

#### Updating the Latent State Sequence c and the Latent Variables z

1: Given the current values  $\mathbf{c}$  and  $\mathbf{z}$ , propose new values  $\mathbf{c}^{new}$  and  $\mathbf{z}^{new}$  according to  $Q(\mathbf{c}^{new}, \mathbf{z}^{new} \mid \mathbf{c}, \mathbf{z}, \boldsymbol{\zeta}) = Q(\mathbf{c}^{new} \mid \mathbf{z}, \boldsymbol{\zeta})Q(\mathbf{z}^{new} \mid \mathbf{c}^{new}, \boldsymbol{\zeta})$ , where  $Q(\mathbf{c} \mid \mathbf{z}, \boldsymbol{\zeta}) \propto \prod_{t} \lambda_{\mathbf{z}_{t}}(c_{t})p(\mathbf{y}_{t} \mid c_{t})$  and  $Q(\mathbf{z} \mid \mathbf{c}, \boldsymbol{\zeta}) = \prod_{j=1}^{q} \widetilde{\pi}_{z_{j,t+j}}^{(j)}(c_{t})$ . Accept the proposed values  $\mathbf{c}^{new}$  and  $\mathbf{z}^{new}$  with probability

$$\min \left\{ \left[ \frac{\prod_{t} \boldsymbol{\lambda}_{\mathbf{z}_{t}^{new}}(c_{t}^{new})}{\prod_{t} \boldsymbol{\lambda}_{\mathbf{z}_{t}}(c_{t})} \frac{\prod_{t} \boldsymbol{\lambda}_{\mathbf{z}_{t}^{new}}(c_{t})}{\prod_{t} \boldsymbol{\lambda}_{\mathbf{z}_{t}}(c_{t}^{new})} \right]^{1/\mathcal{T}(m)}, 1 \right\},$$

with  $\mathcal{T}_0$  and  $\mathcal{T}(m) = \max\{\mathcal{T}_0^{1-m/m_0}, 1\}$  denoting the initial and the current annealing temperature, m the current iteration number and  $m_0$  the iteration number at which the temperature reduces to one.

#### Updating k, the Cluster Mappings $\pi_k$ and the Latent Variables z

Given the current values of  $k_j$  and the current clusters  $\mathcal{C} = \{\mathcal{C}_{j,r} : j = 1, \dots, q, r = 1, \dots, k_j\}$ , do the following for  $j = 1, \dots, q$ .

- 2: If  $k_j < C$ , propose to increase  $k_j$  to  $(k_j + 1)$ . If  $k_j > 1$ , propose to decrease  $k_j$  to  $(k_j 1)$ . For  $1 < k_j < C$ , the moves are proposed with equal probabilities. For  $k_j = 1$ , the increase move is selected with probability 1. For  $k_j = C$ , the decrease move is selected with probability 1.
- 3: If an increase move is proposed, randomly split a cluster into two. If a decrease move is proposed, randomly merge two clusters into a single one.
- 4: Accept the proposed moves with acceptance rates based on the approximated marginal likelihood (24) if  $\sum_{\ell} k_{\ell} > q$ . Set the latent variables  $\mathbf{z}$  at the cluster allocation variables determined by the cluster mappings  $z_{j,t+j} \sim \widetilde{\boldsymbol{\pi}}_{k_j}^{(j)}(c_t)$ .

# Updating the Mixture Weights $\widetilde{\pi}_{\mathbf{k}}$

5: The parameters  $\widetilde{\boldsymbol{\pi}}_{k_j}^{(j)}(w_j)$  are determined by the cluster mappings.

# Updating the Transition Distribution Parameters $\lambda_k$ and $\lambda_0$

6: Sample the parameters  $\boldsymbol{\lambda}_{h_1,\dots,h_q}$  and  $\boldsymbol{\lambda}_0$  as in Algorithm 2.

### Updating the Parameters of the Emission Distribution

7: Sample the parameters  $\psi$  and  $\boldsymbol{\beta}$  as in Algorithm 2.

#### Algorithm 2

#### Updating the Latent State Sequence c

1: Sample the  $c_t$ 's from their multinomial full conditionals

$$p(c_t \mid \boldsymbol{\zeta}) \propto \boldsymbol{\lambda}_{z_{1,t},\dots,z_{q,t}}(c_t) \prod_{j=1}^q \pi_{z_{j,t+j}}^{(j)}(c_t) f(\mathbf{y}_t \mid \boldsymbol{\beta}_{c_t}, \boldsymbol{\psi})).$$

#### Updating the Second Level Latent Variables z

2: Sample the  $z_{j,t}$ 's from their multinomial full conditionals

$$p(z_{j,t} = h \mid \boldsymbol{\zeta}, z_{\ell,t} = h_{\ell}, \ell \neq j) \propto \pi_h^{(j)}(w_{j,t}) \lambda_{h_1,\dots,h_{j-1},h,h_{j+1},\dots,h_q}(c_t).$$

#### Updating the Mixture Weights $\pi_k$

3: Let 
$$n_{j,w_j}(h_j) = \sum_t 1\{w_{j,t} = w_j, z_{j,t} = h_j\}$$
. Sample  $\pi_{k_j}^{(j)}(w_j)$  as

$$\{\pi_1^{(j)}(w_j),\ldots,\pi_{k_j}^{(j)}(w_j)\} \mid \boldsymbol{\zeta} \sim \text{Dir}\{\gamma_j + n_{j,w_j}(1),\ldots,\gamma_j + n_{j,w_j}(k_j)\}.$$

## Updating the Transition Distribution Parameters $\lambda_k$ and $\lambda_0$

4: Let 
$$n_{h_1,\dots,h_q}(c) = \sum_t 1\{z_{1,t} = h_1,\dots,z_{q,t} = h_q, c_t = c\}$$
. Sample the  $\lambda_{h_1,\dots,h_q}$ 's as  $\{\lambda_{h_1,\dots,h_q}(1),\dots,\lambda_{h_1,\dots,h_q}(C)\} \mid \boldsymbol{\zeta} \sim \text{Dir}\{\alpha\lambda_0(1) + n_{h_1,\dots,h_q}(1),\dots,\alpha\lambda_0(C) + n_{h_1,\dots,h_q}(C)\}$ .

5: For  $\ell = 1, \ldots, n_{h_1, \ldots, h_q}(c)$ , sample an auxiliary variable  $x_\ell$  as

$$x_{\ell} \mid \boldsymbol{\zeta} \sim \text{Bernoulli} \left\{ \frac{\alpha \lambda_0(c)}{\ell - 1 + \alpha \lambda_0(c)} \right\}.$$

Set  $m_{h_1,...,h_q}(c) = \sum_{\ell} x_{\ell}$ . 6: Set  $m_0(c) = \sum_{(h_1,...,h_q)} m_{h_1,...,h_q}(c)$ . Sample  $\lambda_0$  as

$$\{\lambda_0(1),\ldots,\lambda_0(C)\} \mid \zeta \sim \text{Dir}\{\alpha_0/C + m_0(1),\ldots,\alpha_0/C + m_0(C)\}.$$

#### Updating the Parameters of the Emission Distribution

7: Sample the global parameters  $\psi$  from their full conditionals

$$p(\boldsymbol{\psi} \mid \boldsymbol{\zeta}) \propto p_0(\boldsymbol{\psi}) \prod_{t=1}^T f(\mathbf{y}_t \mid \boldsymbol{\beta}_{c_t}, \boldsymbol{\psi}).$$

8: Sample the cluster specific parameters  $\boldsymbol{\beta}$  from their full conditionals

$$p(\boldsymbol{\beta}_c \mid \boldsymbol{\zeta}) \propto p_0(\boldsymbol{\beta}_c) \prod_{\{t: c_t = c\}} f(\mathbf{y}_t \mid \boldsymbol{\beta}_c, \boldsymbol{\psi}).$$

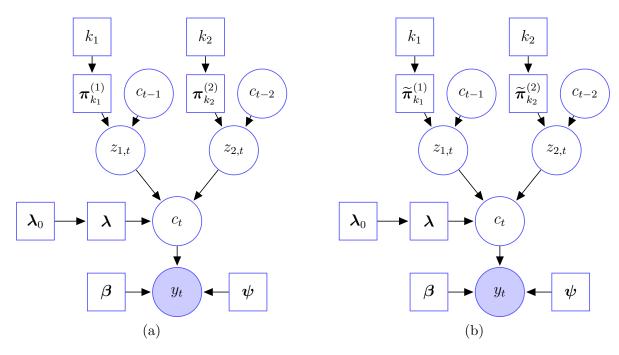


Figure 4: Graphical model depicting the local dependency structure in a second order HMM at time point t. The proposed model (a) and its approximated version (b) that forms the basis of Algorithm 1 described in Section 3. The difference between the two models lies in how the latent variables  $z_{j,t}$ 's are allocated to different latent clusters. The original model implements a soft clustering of  $z_{j,t} \sim \pi_{k_j}^{(j)}(c_{t-j})$  with  $\pi_{k_j}^{(j)}(c_{t-j}) \sim \text{Dir}(\gamma_j, \ldots, \gamma_j)$  for all  $j, c_{t-j}$ . The approximate version implements a hard clustering of  $z_{j,t} \sim \tilde{\pi}_{k_j}^{(j)}(c_{t-j})$  with  $\tilde{\pi}_{h_j}^{(j)}(c_{t-j}) \in \{0,1\}$  for all  $j, h_j$  and  $c_{t-j}$ .

We chose the proposal distributions that mimic their full conditionals and used simulated annealing to facilitate convergence. See Figure 5 and Algorithm 2 for the second stage. The parameters  $\lambda$ ,  $\lambda_0$ ,  $\beta$  and  $\psi$  are updated as in the second stage described in Algorithm 2.

Conditional on  $\mathbf{k}$ , the elements of  $\mathbf{c}$ ,  $\mathbf{z}$  and  $\boldsymbol{\pi}$  all have either multinomial or Dirichlet full conditionals and hence can be straightforwardly updated. Conditional on  $\mathbf{c}$  and  $\mathbf{z}$ , the transition distributions  $\boldsymbol{\lambda}$  and  $\boldsymbol{\lambda}_0$  can be updated adapting existing computational machineries for sampling from the posterior in HDP models. The full conditionals of the parameters characterizing the emission distribution depend on the choice of the parametric family used to model the emission distribution but are usually straightforward to compute, often simplified with conjugate priors on the emission parameters. For fixed  $k_j$ 's, the parameters of the model can thus be easily sampled from the posterior. One such algorithm for finite state space HOHMM, with the auxiliary variable sampler for HDP (Teh et al., 2006) adapted to our setting, is outlined in Algorithm 2. A Chinese restaurant franchise process analog used to derive this algorithm is presented in the Supplementary Materials.

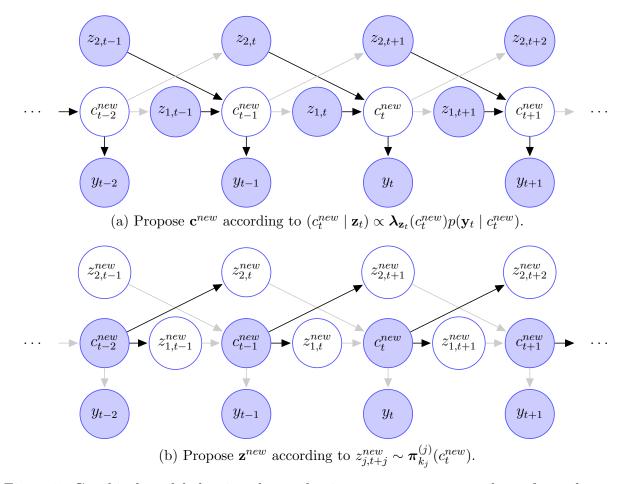


Figure 5: Graphical model showing the mechanisms to propose new values of  $\mathbf{c}$  and  $\mathbf{z}$  in the Metropolis-Hastings step of the approximate sampler described in Section 3. Starting with Figure 2(b), the lighter edges above are ignored in the construction of the proposals.

Additionally, to make the approach even more data adaptive, hyper-priors can be assigned to the prior hyper-parameters  $\alpha$ ,  $\varphi$  etc. and they can also be sampled from the posterior. Priors and full conditionals used to update these hyper-parameters are outlined in the Supplementary Materials.

Specifics of the full conditionals of the emission distribution parameters in steps 7 and 8 of Algorithm 2 depend on the choice of the family and the associated priors. The full conditionals are straightforward for the parametric Normal and Poisson families, and hence are omitted. For the translated normal family, the full conditionals of  $\mu_c$ ,  $\pi_{\eta}$ ,  $\sigma_s^2$  and  $s_t$  are given by

$$(\mu_c|\boldsymbol{\zeta}) \sim \text{Normal}(\mu_{c,T}, \sigma_{c,T}^2),$$
  
 $(\sigma_s^2|\boldsymbol{\zeta}) \sim \text{Inv-Ga}\{a_0 + n_s/2, b_0 + \sum_{t:s_t=s} (y_t - \mu_{c_t} - \eta_s)^2/2\},$   
 $(\boldsymbol{\pi}_{\eta}|\boldsymbol{\zeta}) \sim \text{Dir}(\alpha_{\eta}/S + n_1, \dots, \alpha_{\eta}/S + n_S),$ 

$$p(s_t = s | \boldsymbol{\zeta}) \propto \pi_s \times \text{Normal}(y_t | \mu_{c_t} + \eta_s, \sigma_s^2),$$

where  $n_s = \sum_t 1(s_t = s)$ ,  $\sigma_{c,T}^{-2} = (\sigma_0^{-2} + \sum_{t:c_t=c} \sigma_{s_t}^{-2})$  and  $\mu_{c,T} = \sigma_{c,T}^{-2}(\mu_0 \sigma_0^{-2} + \sum_{t:c_t=c} y_t \sigma_{s_t}^{-2})$ . Without the mean restriction, the posterior full conditional of  $\boldsymbol{\eta}^{S\times 1} = (\eta_1, \dots, \eta_S)^T$  is given by

$$MVN_{S} \left\{ \begin{pmatrix} \eta_{1}^{0} \\ \eta_{2}^{0} \\ \vdots \\ \eta_{S}^{0} \end{pmatrix}, \begin{pmatrix} \sigma_{11}^{0} & 0 & \dots & 0 \\ 0 & \sigma_{22}^{0} & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & \dots & \sigma_{SS}^{0} \end{pmatrix} \right\} \equiv MVN_{S}(\boldsymbol{\eta}^{0}, \boldsymbol{\Sigma}^{0}), \tag{25}$$

where  $\sigma_{ss}^0 = (\sigma_{\eta,0}^{-2} + n_s \sigma_s^{-2})^{-1}$ ,  $\eta_s^0 = \sigma_{ss}^0 \left\{ \sigma_s^{-2} \sum_{t:s_t=s} (y_t - \mu_{c_t}) + \sigma_{\eta,0}^{-2} \mu_{\eta,0} \right\}$ . The posterior full conditional of  $\boldsymbol{\eta}$  under the mean restriction can then be obtained easily by further conditioning (25) by  $\eta_R = \sum_{s=1}^S \pi_s \eta_s = 0$  and is given by

$$(\boldsymbol{\eta} \mid \eta_R = 0, \boldsymbol{\zeta}) \sim \text{MVN}_S \{ \boldsymbol{\eta}^0 - \boldsymbol{\sigma}_{1,R}^0 (\sigma_{R,R}^0)^{-1} \eta_R^0, \boldsymbol{\Sigma}^0 - \boldsymbol{\sigma}_{1,R}^0 (\sigma_{R,R}^0)^{-1} \boldsymbol{\sigma}_{R,1}^0 \},$$
 (26)

where  $\eta_R^0 = \sum_{s=1}^S \pi_s \eta_0 = E(\eta_R)$ ,  $\sigma_{s,S+1} = \pi_s \sigma_{ss}^0 = \text{cov}(\eta_s, \eta_R)$ ,  $\sigma_{R,R}^0 = \sum_{s=1}^S \pi_s^2 \sigma_{ss}^0 = \text{cov}(\eta_R)$ , and  $\boldsymbol{\sigma}_{R,1}^0 = (\sigma_{1,S+1}, \sigma_{2,S+1}, \dots, \sigma_{S,S+1}) = \boldsymbol{\sigma}_{1,R}^{0T}$ . To sample from this singular density, we can first sample from the non-singular distribution of  $\{(\eta_1, \eta_2, \dots, \eta_{S-1})^T \mid \eta_R = 0\}$ , which can also be trivially obtained from (26), and then set  $\eta_S = -\sum_{s=1}^{S-1} \pi_s \eta_s / \pi_S$ .

# 4 Simulation Experiments

We designed simulation experiments to evaluate the performance of our method in a wide range of scenarios.

For the latent state dynamics, we considered the cases (A) [3, {1}], (B) [3, {1,2,3}], (C) [3, {1,2,4}], (D) [3, {1,3,5}], (E) [3, {1,4,8}], where  $[C_0, \{i_1, \ldots, i_r\}]$  means that the latent sequence has  $C_0$  categories and  $\{c_{t-i_1}, \ldots, c_{t-i_r}\}$  are the important lags. In each case, we considered two sample sizes T = 500, 1000. To generate the true transition probability tensors, for each combination of the true lags, we first generated the probability of the first response category as  $f(u_1) = u_1^2/\{u_1^2 + (1-u_1)^2\}$  with  $u_1 \sim \text{Unif}(0,1)$ . The probabilities of the remaining categories are then generated via a stick-breaking type construction as  $f(u_2)\{1-f(u_1)\}$  with  $u_2 \sim \text{Unif}(0,1)$  and so on, until the next to last category (C-1) is reached. The hyper-parameters were set at  $\alpha_0 = 1$ , and  $\gamma_j = 1/C$  for all j. In each case, we set the maximal number of states at C = 10 and the maximal lag at q = 10.

We considered (1) Normal, (2) Poisson and (3) translated mixtures of Normals emission distributions. For the Gaussian case  $f(y \mid c_t = c) = \text{Normal}(y \mid \mu_c, \sigma_c^2)$ , we set  $\mu_c = -2, 0, 2$  for c = 1, 2, 3, respectively, and  $\sigma_c^2 = 0.5^2$  for all c. While the

 $\sigma_c^2$ 's were all equal and could be treated as a global parameter, we allowed the component specific variances to be different in the fitted model. The hyper-parameters of the priors were set at  $\mu_0 = \overline{\mathbf{y}}, \sigma_0^2 = 3\text{var}(\mathbf{y}), a_0 = b_0 = 1$ . For Poisson emission distributions  $f(y \mid c_t = c) = \text{Poi}(y \mid \mu_c)$ , we let  $\mu_c = 1, 8, 15$  for c = 1, 2, 3, respectively. The hyper-parameters of the Gamma prior on  $\mu_c$  were chosen such that  $E(\mu_c) = a_0 b_0 = \overline{\mathbf{y}}, \text{var}(\mu_c) = a_0 b_0^2 = 2\text{var}(\mathbf{y})$ . For translated Gaussian mixture emission distributions  $f(y \mid c_t = c) = \sum_{s=1}^S \pi_s \text{Normal}(y \mid \mu_c + \eta_s, \sigma_s^2)$ , we set  $\mu_c = -4, 0, 4$  for c = 1, 2, 3;  $\pi_s = 0.2, 0.5, 0.3, \eta_s = -2, 0, 1.33$  for s = 1, 2, 3; and  $\sigma_s^2 = 0.5^2$  for all s. As in the case of simpler Gaussian emissions, even though  $\sigma_s^2$ 's were all equal in the true data generating mechanism, they were allowed to be different in the fitted model.

In each case, we initialized the latent states  $\mathbf{c}$  applying a k-means clustering algorithm to  $\mathbf{y}$  with k=C=5 states. With  $k_1=2$  and  $k_j=1$  for  $j=2,\ldots,10$ , initially only the first lag was chosen to be important. The parameters  $\lambda, \lambda_0, \mathbf{z}, \boldsymbol{\pi}$  etc. were then initialized by randomly sampling from the prior generative model. For Normal emission distributions,  $(\mu_c, \sigma_c^2)$ 's were set at the corresponding empirical values; and for Poisson and translated Normal emissions,  $\mu_c$ 's were set at the corresponding empirical means. For translated Normal emissions, the indices  $\mathbf{s}$  and  $(\mu_s, \sigma_s^2)$ 's were likewise set using a k-means algorithm applied to  $(y_t - \mu_{c_t})$  with k=S=5. In each case, the mean parameters associated with the remaining 5 states were spread over the range of  $\mathbf{y}$ . For Normal emissions, the remaining  $\sigma_c^2$ 's were set at  $\operatorname{var}(\mathbf{y})$ . In numerical experiments, we found the results to be very robust to the choice of the prior hyper-parameters and parameter and latent variable initializations.

We coded in MATLAB<sup>1</sup>. For the case (D1) described above, with T = 500 data points, 5,000 MCMC iterations required approximately 30 minutes on an ordinary laptop. For the  $m^{th}$  iteration of the first algorithm, the annealing temperature was set at  $\mathcal{T}(m) = \max\{\mathcal{T}_0^{1-m/m_0}, 1\}$  with  $\mathcal{T}_0 = 1000$  and  $m_0 = 1000$ . In each case, we discarded the first 2000 iterations as burn-in. The remaining samples were thinned by retaining every  $5^{th}$  sample after the burn-in to reduce auto-correlation. The resulting samples showed good mixing and convergence behavior in diagnostic checks and produced stable estimates of the parameters of interest.

We evaluated the performance of our proposed model and the HDP-HMM in estimating one, two and three-step ahead predictive densities. For an HOHMM of maximal order q, the r-step ahead predictive density is given by

$$f_{pred,T+r}(y|\mathbf{y}_{1:T}) = E_{P(\boldsymbol{\zeta},\mathbf{c}|\mathbf{y}_{1:T})}p(y\mid\mathbf{c},\boldsymbol{\zeta}) = \int p(y\mid\mathbf{c},\boldsymbol{\zeta})dP(\boldsymbol{\zeta},\mathbf{c}\mid\mathbf{y}_{1:T})$$

<sup>&</sup>lt;sup>1</sup>Codes implementing our method will be included as part of the Supplementary Materials once the paper is accepted for publication.

Thurs	Sample Size	Median ISE $\times 100$							
True Dynamics		HDP-HMM			CTF-HOHMM				
		One	Two	Three	One	Two	Three		
Normal Emission Distribution									
(A) 3, {1}	500	0.37	0.20	0.23	0.37(0.33)	0.19 (0.17)	0.20 (0.19)		
	1000	0.16	0.13	0.11	0.11(0.11)	0.11(0.11)	0.09 (0.10)		
(B) 3, {1, 2, 3}	500	10.96	8.46	4.21	0.82	0.67	0.65		
	1000	4.47	5.04	3.26	0.32	0.31	0.20		
(C) $3, \{1, 2, 4\}$	500	11.39	8.46	6.66	0.88	0.80	0.65		
	1000	9.52	6.51	5.98	0.31	0.30	0.29		
(D) 3, {1, 3, 5}	500	10.91	11.14	9.38	0.71	0.69	0.57		
	1000	15.60	11.24	9.21	0.30	0.35	0.32		
Poisson Emission Distribution									
(A) 2 (1)	500	0.15	0.22	0.10	0.09(0.07)	0.07 (0.05)	0.05 (0.04)		
(A) $3, \{1\}$	1000	0.30	0.26	0.17	0.09(0.04)	0.05(0.04)	0.03 (0.03)		
(B) 3, {1, 2, 3}	500	3.02	1.92	1.77	1.17	0.61	0.65		
	1000	1.99	1.65	1.32	0.32	0.14	0.09		
(C) $3, \{1, 2, 4\}$	500	2.58	1.75	1.62	0.57	0.47	0.45		
	1000	3.61	2.41	1.50	0.72	0.51	0.32		
(D) 3, {1, 3, 5}	500	4.72	1.71	2.71	2.41	1.26	0.78		
	1000	5.08	2.37	3.24	1.03	0.76	0.53		
Translated Mixture Normal Emission Distribution									
(A) 3, {1}	500	0.35	0.27	0.27	0.39(0.35)	0.29 (0.29)	0.27 (0.27)		
	1000	0.21	0.15	0.11	0.23(0.22)	0.20(0.15)	0.12 (0.12)		
(B) 3, {1, 2, 3}	500	5.39	3.40	2.99	1.25	1.09	0.85		
	1000	4.09	3.84	2.60	0.53	0.47	0.31		
(C) $3, \{1, 2, 4\}$	500	6.55	3.88	3.37	1.44	1.401	1.06		
	1000	3.51	3.00	2.84	0.58	0.46	0.43		
(D) $3, \{1, 3, 5\}$	500	7.80	5.15	2.81	3.06	1.76	1.21		
	1000	7.20	3.77	3.19	0.53	0.59	0.47		

Table 1: Median ISEs in estimating one, two and three step ahead predictive densities for the conditional tensor factorization (CTF) based HOHMM compared with the HDP-HMM. In the first column,  $C_0, \{i_1, \ldots, i_r\}$  means that the latent sequence truly has  $C_0$  categories and  $\{c_{t-i_1}, \ldots, c_{t-i_r}\}$  are the true important lags. In the rows corresponding to the first order case (A) 3,  $\{1\}$ , the numbers within parenthesis in the CTF-HOHMM columns show the estimated median MISEs with the maximal order set at q=1. In all other cases, q=5. See Section 4 for additional details.

$$= E_{P(\boldsymbol{\zeta}, \mathbf{c} | \mathbf{y}_{1:T})} \left[ \sum_{c_{T+r}} \sum_{c_{T+r-1}} \cdots \sum_{c_{T+1}} f(y \mid c_{T+r}, \boldsymbol{\zeta}) p(c_{T+r} \mid \mathbf{c}_{(T+r-q):(T+r-1)}, \boldsymbol{\zeta}) \dots p(c_{T+1} \mid \mathbf{c}_{(T+1-q):T}, \boldsymbol{\zeta}) \right].$$

Based on M samples  $\{(\mathbf{c}_{1:T}^{(m)}, \boldsymbol{\zeta}^{(m)})\}_{m=1}^{M}$  drawn from the posterior,  $f_{pred,T+r}(y|\mathbf{y}_{1:T})$  can be estimated as

$$\widehat{f}_{pred,T+r}(y|\mathbf{y}_{1:T}) = M^{-1} \sum_{m=1}^{M} \sum_{c_{T+r}} \sum_{c_{T+r-1}} \cdots \sum_{c_{T+1}} f(y \mid c_{T+r}, \boldsymbol{\zeta}^{(m)})$$

True	Sample	Median Hamming Distance $\times 100$						
Dynamics	Size	HDP-HMM	CTF-HOHMM					
Normal Emission Distribution								
(A) 2 (1)	500	3.77	2.56(2.23)					
(A) $3, \{1\}$	1000	2.57	2.10(2.00)					
(D) 2 [1 9 2]	500	19.02	2.43					
(B) $3, \{1, 2, 3\}$	1000	16.37	2.24					
(C) $3, \{1, 2, 4\}$	500	20.46	2.50					
	1000	19.82	2.19					
(D) 2 (1 2 E)	500	16.45	2.33					
(D) $3, \{1, 3, 5\}$	1000	12.01	2.30					
Poisson Emission Distribution								
(A) p. (1)	500	11.01	8.72 (8.10)					
(A) $3, \{1\}$	1000	12.78	8.12 (8.01)					
(D) 2 (1 0 2)	500	23.45	12.66					
(B) $3, \{1, 2, 3\}$	1000	21.50	8.92					
(C) 2 (1 2 4)	500	23.29	10.95					
(C) $3, \{1, 2, 4\}$	1000	22.11	8.91					
(D) 3, {1, 3, 5}	500	21.35	14.26					
(D) 3, {1, 3, 3}	1000	21.03	10.96					
Translated Mixture Normal Emission Distribution								
(A) 2 (1)	500	4.83	7.31 (5.80)					
(A) $3, \{1\}$	1000	4.38	6.81 (5.58)					
(B) 3, {1, 2, 3}	500	19.21	7.02					
$(D)$ 3, $\{1, 2, 3\}$	1000	24.01	6.92					
(C) $3, \{1, 2, 4\}$	500	17.36	10.38					
(0) 3, (1, 2, 4)	1000	20.59	5.93					
(D) 3, {1, 3, 5}	500	14.09	15.88					
(12) 3, \(\gamma\),	1000	13.81	5.32					

Table 2: Median Normalized Hamming distances between the true and the estimated state sequences for the conditional tensor factorization (CTF) based HOHMM and the HDP-HMM. In the first column,  $C_0$ ,  $\{i_1, \ldots, i_r\}$  means that the latent sequence truly has  $C_0$  categories and  $\{c_{t-i_1}, \ldots, c_{t-i_r}\}$  are the true important lags. In the rows corresponding to the first order case (A) 3,  $\{1\}$ , the numbers within parenthesis in the CTF-HOHMM columns show the estimated median MISEs with the maximal order set at q=1. In all other cases, q=5. See Section 4 for additional details.

$$p(c_{T+r} \mid \mathbf{c}_{(T+r-q):(T+r-1)}^{(m)}, \boldsymbol{\zeta}^{(m)}) \dots p(c_{T+1} \mid \mathbf{c}_{(T+1-q):T}^{(m)}, \boldsymbol{\zeta}^{(m)}),$$

where  $\mathbf{c}_{(T+r-q):(T+r-1)}^{(m)} = (c_{T+r-q}^{(m)}, c_{T+r-q+1}^{(m)}, \dots, c_{T}^{(m)}, c_{T+1}, \dots, c_{T+r-1})$  for all (r, m). The corresponding true density is given by

$$f_{pred,T+r,0}(y) = \sum_{c_{T+r,0}} \sum_{c_{T+r-1,0}} \cdots \sum_{c_{T+1,0}} f_0(y \mid c_{T+r,0})$$

$$p_0(c_{T+r,0} \mid \mathbf{c}_{(T+r-q):(T+r-1),0}) \cdots p_0(c_{T+1,0} \mid \mathbf{c}_{(T+1-q):T,0}),$$

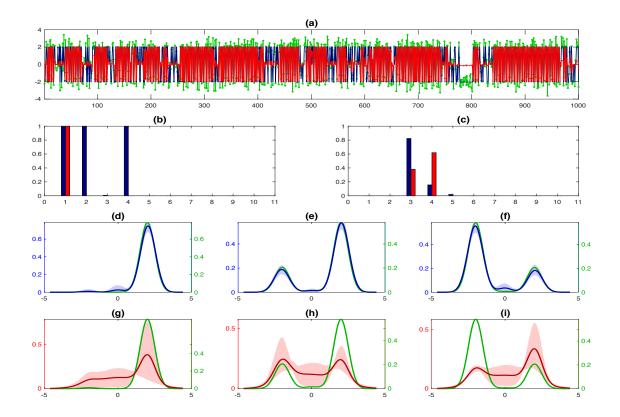


Figure 6: Results for a synthetic dataset in the (D1) case described in Section 4 (transition dynamics 3{1,2,4} and Normal emission densities): CTF-HOHMM in blue and HDP-HMM in red (a) posterior means super-imposed over the observed time series in green; (b) the inclusion probabilities of different lags; (c) the distribution of different number of states; (d), (e), (f) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the CTF-HOHMM; (g), (h), (i) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the HDP-HMM. The true predictive densities are shown in green.

where  $p_0$  and  $f_0$  are generics for the true transition and emission distributions, respectively, with associated true parameters implicitly understood and  $\mathbf{c}_0 = \mathbf{c}_{1:T,0}$  denoting the true values of the latent sequence  $\mathbf{c}$ . For continuous emission distributions, the integrated squared error (ISE) in estimating  $f_{pred,T+r,0}(y)$  is estimated by  $\sum_{i=1}^{N} \{f_{pred,T+r,0}(y_i^{\Delta}) - \widehat{f}_{pred,T+r}(y_i^{\Delta} \mid \mathbf{y}_{1:T})\}^2 \Delta_i$ , where  $\{y_i^{\Delta}\}_{i=0}^{N}$  are a set of grid points on the range of y and  $\Delta_i = (y_i^{\Delta} - y_{i-1}^{\Delta})$  for all i. For Poisson emission distribution, the ISE is estimated as  $\sum_{i=\max\{0,\min\mathbf{y}-1\}}^{\max\mathbf{y}+1} \{f_{pred,T+r,0}(i) - \widehat{f}_{pred,T+r}(i \mid \mathbf{y}_{1:T})\}^2$ .

We also evaluated the Hamming distance between the true and the estimated state sequences by the proposed HOHMM and the HDP-HMM. To calculate the Hamming distance, we used the Munkres algorithm (Munkres, 1957), mapping the indices of the estimated state sequence to the set of indices that maximize the overlap with the true sequence.

The performances in estimating the one, two and three step ahead predictive

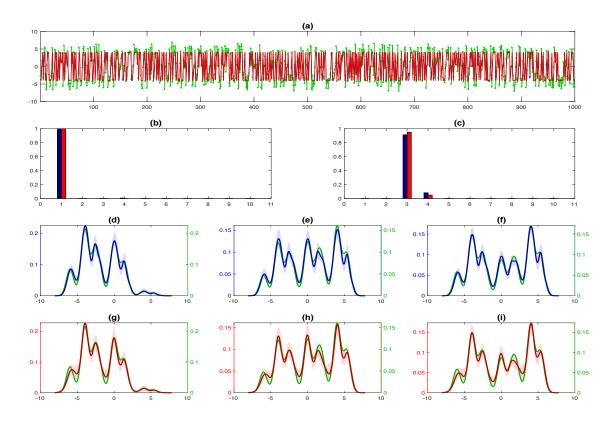


Figure 7: Results for a synthetic dataset in the (A3) case described in Section 4 (transition dynamics 3{1} and translated Normal emission densities): CTF-HOHMM in blue and HDP-HMM in red (a) posterior means super-imposed over the observed time series in green; (b) the inclusion probabilities of different lags; (c) the distribution of different number of states; (d), (e), (f) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the CTF-HOHMM; (g), (h), (i) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the HDP-HMM. The true predictive densities are shown in green.

densities and in clustering the observations  $y_t$  are summarized in Tables 1 and 2, respectively. The reported results are based on 100 simulated datasets in each case. The proposed approach vastly outperformed the HDP-HMM in the higher order cases and remarkably also in the first order parametric cases. Figure 6 summarizes the results for the data set corresponding to the median ISE in estimating the one-step ahead predictive density for the HOHMM in the (D1) case with T=1000. Panel (a) in Figure 6 suggests that the CTF-HOHMM provides a better fit to local variations in the dataset. The improvements in higher order cases are explained by the HDP-HMM's restrictive first order assumption. The proposed method, on the other hand, not just accommodates higher order lags, but also effectively eliminates the unnecessary ones, while also characterizing the dynamics using efficient sparse representations. As explained in Section 2.6, the improvements in the first order cases can be attributed to this ability to effectively eliminate the unnecessary lags in correctly identifying

the true first order dynamics and then sparsely characterizing the dynamics using soft allocation of the latent states, enabling better data compression than the hard allocation model implemented in HDP-HMM.

The remarkable efficiency of the proposed HOHMM method even when the maximal lag is set at large conservative values is also seen from comparisons of the results when the maximal lag was set at 10 with the results reported in parentheses in Tables 1 and 2 that were produced by its first order restriction by prefixing the maximal lag at one. See Figure 7 that summarizes the results for the data set corresponding to the median ISE for the one-step ahead predictive density in the (A3) case with T = 1000. Table 1 shows that with increase in the prediction step the performance in estimating the predictive density improved for both the HDP-HMM and the proposed HOHMM. As the prediction step increases, the true and the estimated predictive densities approach the corresponding stationary distributions which are invariant to  $c_{T-i_1+1}, \ldots, c_{T-i_r+1}$  and the error stabilizes. Improved estimation of the latent states, as evident from the estimated Hamming distances, can likewise be explained by the aforementioned novel aspects of the proposed HOHMM.

# 5 Applications

In this section, we discuss results of the proposed CTF-HOHMM applied to a few real datasets. The datasets discussed here are all available publicly. In each case, we compare with HDP-HMM. Unless otherwise mentioned, each model is allowed a maximum of C=10 states; the HOHMM was allowed a maximal lag of q=10; and the model parameters were all initialized as in the simulation experiments.

Old Faithful Geyser Data: We first consider the Geyser dataset, accompanying McDonald and Zuchhini (1997) and also available from the MASS package in R. The dataset comprises 299 sequential observations on eruption and waiting times (in minutes) of the Old Faithful geyser in Yellowstone National Park in the USA collected continually from August 1 to August 15, 1985. We focus here on modeling duration times using HMMs with Normal emission distributions. Empirical explorations of the dataset earlier in Section 3 in Azzalini and Bowman (1990) had suggested second order dynamics.

Figure 8 summarizes the results. Results produced by HDP-HMM and CTF-HOHMM are in general agreement, both models suggesting a three state dynamics. The results returned by HOHMM, however, suggest a second order HMM to provide the best fit, consistent with Azzalini and Bowman (1990).

MIT Heart Data: Next, we consider MIT heart data, a collection of 4 time series. The first two series contains 1800 evenly-spaced measurements of instantaneous heart rate from different subjects performing similar activities. The measurements (in units of beats per minute) occur at 0.5 second intervals over a 15 minute interval.

Figure 9 summarizes the results for the series 1 dataset obtained by the CTF-

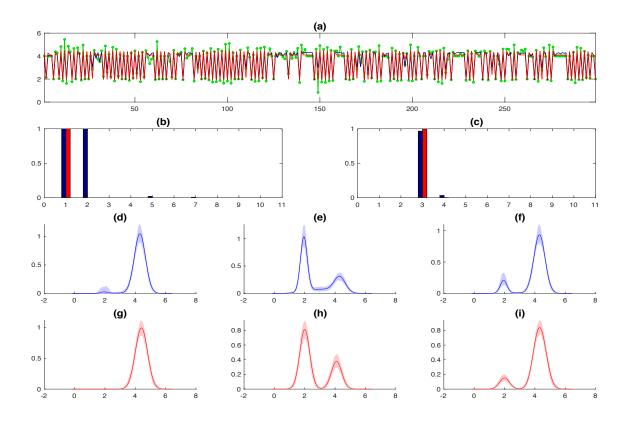


Figure 8: Results for the Geyser dataset: CTF-HOHMM in blue and HDP-HMM in red (a) posterior means super-imposed over the observed time series in green; (b) the inclusion probabilities of different lags; (c) the distribution of different number of states; (d), (e), (f) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the CTF-HOHMM; (g), (h), (i) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the HDP-HMM.

HOHMM and the HDP-HMM with Normal emission distributions. The HDP-HMM results show uncertainty around the number of underlying latent states, suggesting a mixture of 5 and 6 latent states. The CTF-HOHMM results suggest second order dependencies. Like the HDP-HMM, CTF-HOHMM also accommodates uncertainty in the number of states, suggesting however a mixture of 8 and 9 states. Panel (a) in Figure 9 suggests that the CTF-HOHMM provides a better fit to local variations in the dataset. The predictive densities estimated by the two methods also look substantially different.

The series 2 in the MIT heart dataset shows strong signs of irregular periodicity. HMMs are not suitable for modeling periodicity without additional modifications. We have thus not pursued modeling series 2. The series 3 and 4 were recorded in the same way but contain 950 measurements each, corresponding to 7 minutes and 55 seconds of data in each case. CTF-HOHMM applied to these two datasets suggests first order dependencies in both cases. Results produced by HDP-HMM and CTF-HOHMM,

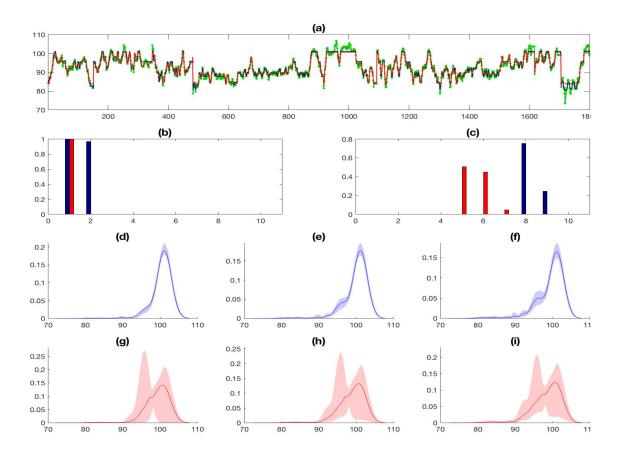


Figure 9: Results for the MIT heart dataset: CTF-HOHMM in blue and HDP-HMM in red (a) posterior means super-imposed over the observed time series in green; (b) the inclusion probabilities of different lags; (c) the distribution of different number of states; (d), (e), (f) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the CTF-HOHMM; (g), (h), (i) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the HDP-HMM.

not presented here, were very similar for these two series.

**E.coli Data:** Next, we consider the E.coli data set available from tscount package in R. This dataset comprises weekly counts of reported disease cases caused by Escherichia coli in the state of North RhineWestphalia (Germany) from January 2001 to May 2013.

Figure 10 summarizes the results obtained by CTF-HOHMM and HDP-HMM with Poisson emission distributions. The HDP-HMM results suggests 5 latent states. The CTF-HOHMM results suggests a first order dynamics but 7 latent states. Panel (a) in Figure 10 suggests that the CTF-HOHMM provides a better fit to local variations in the data. The one, two and three steps ahead predictive densities, however, look similar.

Coriell aCGH Data: Array comparative genomic hybridization (aCGH) stud-

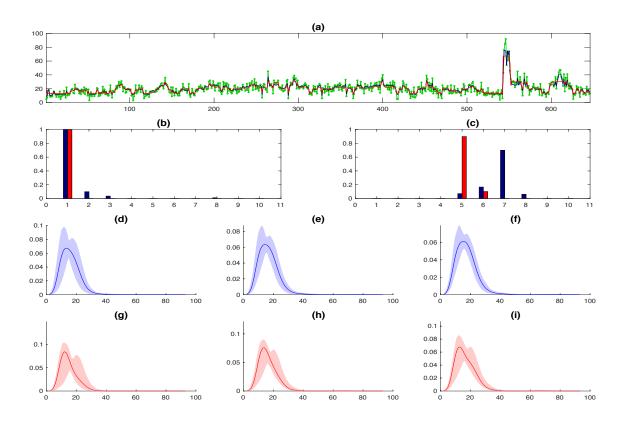


Figure 10: Results for the E.coli dataset: CTF-HOHMM in blue and HDP-HMM in red (a) posterior means super-imposed over the observed time series in green; (b) the inclusion probabilities of different lags; (c) the distribution of different number of states; (d), (e), (f) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the CTF-HOHMM; (g), (h), (i) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the HDP-HMM.

ies are used to investigate the identification of DNA polymorphisms - deletions or sequences deviations, and duplications. The measurement at the  $t^{th}$  location on the chromosome typically represents  $\log_2$  ratio of copy numbers in the test genome to that in a reference genome. A zero value thus indicates the same copy number as that in the reference genome, positive values indicate gains or amplifications, whereas negative observations represent deletions.

First and higher order HMMs have been used in the literature for modeling aCGH data. Fridlyand et al. (2004) used a first order HMM to segment aCGH data into sets with the same underlying copy number; Guha et al. (2008) used a first order HMM with four latent copy number states with Gaussian emission densities; Seifert et al. (2012) used a tree based higher order HMM with three states with Gaussian emission densities; Yau et al. (2011) used a first order HMM with three states but a flexible infinite component translated-mixture of Normals as the emission distribution. The use of a flexible family of emission densities made this approach robust to the presence

of outliers, skewness or heavy tails in the error process.

We consider the Coriell aCGH dataset from the DNAcopy package in Bioconductor, originally presented in Snijders et al. (2001). The data correspond to two array CGH studies of fibroblast cell strains. We chose the study GM05296 that comprised copy number ratios at 2271 consecutive genomic locations. We model the dataset using HDP-HMM and the proposed HOHMM with C=3 states. The unknown state specific means  $\mu_c$ 's are allowed to vary according to Normal( $\mu_{c,0}, \sigma_{c,0}^2$ ) hyper-priors with  $\mu_{c,0} = -0.5, 0.0, 0.5$  and  $\sigma_{c,0} = 1/6, 10^{-5}, 1/6$  for c=1,2,3, allowing  $\mu_c$ 's to vary over [-1,1] essentially across disjoint intervals, thus ensuring identifiability of these states. As in Yau et al. (2011), we also use a translated-mixture of Normals as our emission distribution with S=5 local components. Experiments with larger values of S did not result in any practical difference in the results.

Figure 11 summarizes the results for the Coriell aCGH dataset obtained by CTF-HOHMM and HDP-HMM. The CTF-HOHMM results suggest higher order dependence with the first three lags being the important ones. This is reflective of the fact that copy number variations usually occur in clusters of adjacent locations. Panel (a) in Figure 11 suggests that the CTF-HOHMM provides a better fit to local variations in the data, better capturing focal aberrations (Fridlyand et al., 2004) due to alterations in very narrow regions. The predictive densities estimated by the two methods also look quite different. For example, for one step ahead prediction, the HDP-HMM basically predicts a Normal copy number state. The CTF-HOHMM, on the other hand, assigns equal probabilities to having either a Normal copy number or an increased copy number. It takes in account not just the immediately preceding location, which had a Normal copy number state, but also the variation in a few preceding locations which had amplified copy numbers.

# 6 Discussion

We proposed a flexible yet parsimonious nonparametric Bayesian approach to higher order hidden Markov models that allows automated identification of the important lags. The celebrated HDP-HMM is obtained as a special case when the order is restricted to one and the soft sharing feature of the model is turned off. In simulation experiments, our method vastly out-performed the HDP-HMM in higher order settings. Remarkably, the improvements were also substantial in the first order cases which may be attributed to greater levels of data compression achievable by the proposed model.

The focus of this paper has been on higher order homogeneous HMMs, but the proposed methodology can be easily extended to nonhomogeneous cases where the transition dynamics and the emission distributions are allowed to be influenced by exogenous predictors. We are also pursuing the development of faster algorithms for online inference in HOHMMs that scale better with larger datasets. Additional

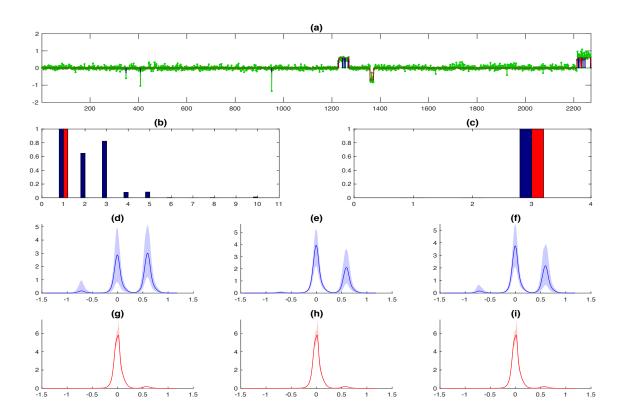


Figure 11: Results for the aCGH dataset: CTF-HOHMM in blue and HDP-HMM in red (a) posterior means super-imposed over the observed time series in green; (b) the inclusion probabilities of different lags; (c) the distribution of different number of states; (d), (e), (f) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the CTF-HOHMM; (g), (h), (i) estimated one, two and three steps ahead predictive densities, respectively, and their 90% credible intervals by the HDP-HMM.

important directions of ongoing research include extensions to other discrete state space dynamical systems, and models for spatial and spatio-temporal datasets.

# References

Albert, J. and Chib, S. (1993). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Business and Economic Statistics*, 11, 1–15.

Alexandrovich, G., Holzmann, H., and Leister, A. (2016). Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, **103**, 423–434.

Azzalini, A. and Bowman, A. H. (1990). A look at some data on the Old Faithful geyser. *Journal of the Royal Statistical Society: Series C*, **39**, 357–365.

- Bae, K., Mallick, B. K., and Elsik, C. G. (2005). Prediction of protein interdomain linker regions by a hidden Markov model. *Bioinformatics*, **21**, 2264–2270.
- Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Verlag, Berlin.
- De Lathauwer, L., De Moore, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. SIAM Journal on Matrix Analysis and Applications, 21, 1253–1278.
- Ferguson, T. F. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209–230.
- Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, 5, 1020–1056.
- Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004). Hidden Markov models approach to the analysis of array CGH data. *Journal of multivariate analysis*, **90**, 132–153.
- Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, **96**, 194–209.
- Frühwirth-Schnatter, S. (2006). Finite mixture and Markov switching models. Springer, New York.
- Gassiat, E. and Rousseau, J. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli*, **20**, 2039–2075.
- Gassiat, E., Cleynen, A., and Robin, S. (2015). Inference in finite state space non-parametric hidden Markov models and applications. *Statistics and Computing*, pages 10.1007/s11222-014-9523-8.
- George, E. and McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–373.
- Guha, S., Li, Y., and Neuberg, D. (2008). Bayesian hidden Markov modeling of array CGH data. *Journal of the American Statistical Association*, **103**, 485–497.
- Hamilton, J. E. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, **45**, 39–70.
- Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, **30**, 158–182.
- Jääskinen, V., Xiong, J., Corander, J., and Koski, T. (2014). Sparse Markov chains for sequence data. *Scandinavian Journal of Statistics*, **41**, 639–655.

- Langrock, R., Kneib, T., Sohn, A., and DeRuiter, S. L. (2015). Nonparametric inference in hidden markov models using p-splines. *Biometrics*, **71**, 520–528.
- Lennox, K. P., Dahl, D. B., Day, R., and Tsai, W. (2010). A Dirichlet process mixture of hidden Markov models for protein structure prediction. *Annals of Applied Statistics*, 4, 916–942.
- Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. Stochastic Processes and their Applications, 40, 127–143.
- Leroux, B. G. and Puterman, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, pages 545–558.
- McDonald, S. and Zuchhini, W. (1997). Hidden Markov and other models for discretevalued time series. Chapman & Hall, London.
- Munkres, J. (1957). Algorithms for the assignment and transportation problems. Journal of the Society of Industrial and Applied Mathematics, 5, 32–38.
- Quick, N. J., Isojunno, S., Sadykova, D., Bowers, M., Nowacek, D. P., and Read, A. J. (2017). Hidden Markov models reveal complexity in the diving behaviour of short-finned pilot whales. *Scientific Reports*, 7, 45765.
- Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE*, **77**, 257–286.
- Robert, C. P., Ryden, T., and Titterington, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society: Series B*, **62**, 57–75.
- Sarkar, A. and Dunson, D. B. (2016). Bayesian nonparametric modeling of higher-order Markov chains. *Journal of the American Statistical Association*, **111**, 1791–1803.
- Sclove, S. L. (1983). Time-series segmentation: A model and a method. *Information Sciences*, **29**, 7–25.
- Scott, S. L. (2002). Bayesian methods for hidden Markov models recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**, 337–351.
- Seifert, M., Gohr, A., Strickert, M., and Grosse, I. (2012). Parsimonious higher-order hidden Markov models for improved array-CGH analysis with applications to Arabidopsis thaliana. *PLoS Computational Biology*, 8, 1–15.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639–650.
- Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., et al. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, 29, 263–264.

- Teh, Y. W., Jordan, M. I., J., B. M., and M., B. D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, **101**, 1566–1581.
- Thede, S. M. and Harper, M. P. (1999). A second-order hidden markov model for part-of-speech tagging. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 175–182. Association for Computational Linguistics.
- Titsias, M. K., Holmes, C. C., and Yau, C. (2016). Statistical inference in hidden Markov models using k-segment constraints. *Journal of the American Statistical Association*, **111**, 200–215.
- Tucker, L. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrica*, **31**, 273–282.
- Vernet, E. (2015a). Nonparametric hidden Markov models with finite state space: posterior concentration rates. arXiv preprint.
- Vernet, E. (2015b). Posterior consistency for nonparametric hidden Markov models with finite state space. *Electronic Journal of Statistics*, **9**, 717–752.
- Wang, P. and Puterman, M. L. (1999). Markov Poisson regression models for discrete time series. part 1: Methodology. *Journal of Applied Statistics*, **26**, 855–869.
- Yang, Y. and Dunson, D. B. (2016). Bayesian conditional tensor factorization for high-dimensional classification. *Journal of the American Statistical Association*, 111, 656–669.
- Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011). Bayesian nonparametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society, Series B*, **73**, 37–57.

# Supplementary Materials for **Bayesian Higher Order Hidden Markov**Models

Abhra Sarkar
Department of Statistics and Data Sciences,
University of Texas at Austin,
2317 Speedway D9800, Austin, TX 78712-1823, USA
abhra.sarkar@utexas.edu
and
David B. Dunson
Department of Statistical Science,
Duke University,
Box 90251, Durham NC 27708-0251
dunson@duke.edu

# Summary

The supplementary materials discuss details of the theoretical results on model identifiability and posterior convergence from Section 2.2 and Section 2.4 of the main paper and also provide some additional results. The supplementary document also presents details of the Chinese restaurant franchise process analog used to design the MCMC sampling algorithm described in Section 3 of the main paper. The document concludes with algorithms for sampling some of the prior hyper-parameters.

## S.1 Proof of Lemma 1

We refer to an HOHMM with transition probability tensor  $\mathbf{p}$ , initial and stationary distribution  $\boldsymbol{\pi}$ , and emission distributions  $\mathbf{f}$  by HOHMM( $\boldsymbol{\pi}, \mathbf{p}, \mathbf{f}$ ). The joint distribution of  $\mathbf{y}_{1:T_0}$ ,  $T_0 = (2C^q + 1)q$ , under HOHMM( $\boldsymbol{\pi}, \mathbf{p}, \mathbf{f}$ ) is then given by

$$p_{\mathbf{p},\mathbf{f}}(\mathbf{y}_{1:T_0}) = \sum_{c_1,\dots,c_{T_0}} \pi_{c_1,\dots,c_q} p_{c_{q+1}|c_1,\dots,c_q} \dots p_{c_{T_0}|c_{T_0-q},\dots,c_{T_0-1}} f_{c_1} \dots f_{c_{T_0}}$$

$$= \sum_{c_1,\dots,c_{T_0}} \pi_{c_1,\dots,c_q} p_{c_{q+1},\dots,c_{2q}|c_1,\dots,c_q} \dots p_{c_{T_0-q},\dots,c_{T_0}|c_{T_0-2q},\dots,c_{T_0-q-1}} (f_{c_1} \dots f_{c_q}) \dots (f_{c_{T_0-q}} \dots f_{c_{T_0}}).$$

The elements of  $\widetilde{P}^q$  are the probabilities  $p_{c_{q+1},\dots,c_{2q}|c_1,\dots,c_q} = p\{(c_{q+1},\dots,c_{2q}) \mid (c_1,\dots,c_q)\}$  of moving from  $(c_1,\dots,c_q)$  to  $(c_{q+1},\dots,c_{2q})$  in q steps by a Markov chain with transition probability matrix  $\widetilde{P}$ . Also, since  $f_c, c \in \mathcal{C}$ , are all distinct in  $\mathcal{Y}$ , so are their ordered products  $f_{c_1}f_{c_2}\cdots f_{c_q}$  in  $\mathcal{Y}^q$ . Additionally, since  $\widetilde{P}$  is of full rank, so is  $\widetilde{P}^q$ . Straightforward application of Theorem 1 from Alexandrovich et al. (2016) then establishes separate nonparametric identifiability of  $\widetilde{P}^q$  and  $\{f_{c_1}f_{c_2}\cdots f_{c_q}: c_j \in \mathcal{C}, j=1,\dots,q\}$  up to label swapping of the states from the joint distribution of  $(2C^q+1)q$  consecutive observations.

Let  $\text{HOHMM}(\boldsymbol{\pi}^{\star}, \mathbf{p}^{\star}, \mathbf{f}^{\star})$  be an HOHMM such that  $p_{\mathbf{p}, \mathbf{f}}(\mathbf{y}_{1:T_0}) = p_{\mathbf{p}^{\star}, \mathbf{f}^{\star}}(\mathbf{y}_{1:T_0})$ . There thus exists a permutation  $\sigma$  such that

$$f_{c_1}^{\star} \dots f_{c_q}^{\star} = f_{\sigma_1(c_1,\dots,c_q)} \dots f_{\sigma_q(c_1,\dots,c_q)} \quad \text{and}$$

$$p_{c_{q+1},\dots,c_{2q}|c_1,\dots,c_q}^{\star} = p_{\sigma_1(c_{q+1},\dots,c_{2q}),\dots,\sigma_q(c_{q+1},\dots,c_{2q})|\sigma_1(c_1,\dots,c_q),\dots,\sigma_q(c_1,\dots,c_q)},$$

where  $\sigma_j(c_1,\ldots,c_q)$  is the  $j^{th}$  element of  $\sigma(c_1,\ldots,c_q)$ .

Since  $f_c^*$  and  $f_c$  are probability distributions, for any pair  $(j, c_j)$  we have  $f_{c_j}^* = f_{\sigma_j(c_1, \dots, c_q)}$  for all  $c_\ell \in \mathcal{C}, \ell \neq j$ , which implies  $\sigma_j(c_1, \dots, c_q) = \sigma(c_j)$  for all  $c_j \in \mathcal{C}, j = 1, \dots, q$ , and hence  $f_c^* = f_{\sigma(c)}$  for all  $c \in \mathcal{C}$ . This also implies that

$$p_{c_{q+1},\dots,c_{2q}|c_1,\dots,c_q}^{\star} = p_{\sigma(c_{q+1}),\dots,\sigma(c_{2q})|\sigma(c_1),\dots,\sigma(c_q)}$$
 for all  $(c_{q+1},\dots,c_{2q})$ .

Summing across  $c_{q+2}, \ldots, c_{2q}$  then gives

$$p_{c_{q+1}|c_1,\dots,c_q}^{\star} = p_{\sigma(c_{q+1})|\sigma(c_1),\dots,\sigma(c_q)}$$
 for all  $(c_1,\dots,c_q,c_{q+1})$ .

Finally, since the stationary distribution  $\pi$  is uniquely determined by  $\widetilde{P}$ , we have for all  $(c_1, \ldots, c_q)$ ,  $\pi_{c_1, \ldots, c_q}^{\star} = \pi_{\sigma(c_1), \ldots, \sigma(c_q)}$ . This concludes the proof of Lemma 1.

### S.2 Proof of Theorem 1 and Some Additional Results

Using the first order representation of HOHMMs in blocks of size q considered in Section 2.2, Theorem 1 can be proved with  $\underline{p}$  replaced by  $\underline{p}^q$  for values of  $\ell$  that are multiples of q by straightforwardly adapting the proofs of the results for first order HMMs in Vernet (2015b). For general values of  $\ell \geq q$ , we instead combine the first q observations into a special  $0^{th}$  initial state  $(c_1, \ldots, c_q)$  and then consider single step transitions as

$$(c_1, \dots, c_q) \rightarrow (c_2, \dots, c_{q+1}) \rightarrow (c_3, \dots, c_{q+2}) \rightarrow \cdots$$

$$\downarrow \qquad \qquad \downarrow \qquad \qquad \downarrow$$

$$(\mathbf{y}_1, \dots, \mathbf{y}_q) \qquad \qquad \mathbf{y}_{q+1} \qquad \qquad \mathbf{y}_{q+2} \qquad \cdots$$

Using the above representation and adapting to Vernet (2015b), the following additional results can also be easily established. Theorem 2 below establishes consistency the  $\ell^{th}$  order marginals in the weak topology (Ghosh and Ramamoorthi, 2003, page 12). Theorem 3 establishes separate consistency of the transition probabilities and the emission distributions, the latter in the weak topology.

**Theorem 2.** Under Assumptions 1, for all weak neighborhood  $U_W$  of  $f_{1:\ell,0}^{\star}$ ,  $\ell \geq q$ ,

$$P_{\pmb{\theta}_0} \left\{ \lim_{T \to \infty} \Pi(U_W \mid \mathbf{y}_{1:T}) = 1 \right\} = 1.$$

Theorem 3. Let  $\boldsymbol{\theta}_0 = (\mathbf{p}_0, \mathbf{f}_0)$  characterize a stationary HOHMM such that  $\widetilde{P}_0$ , the first order representation of  $\mathbf{p}_0$ , is full rank and the elements of  $\mathbf{f}_0 = (f_{1,0}, \dots, f_{C,0})^T$  are linearly independent. Also, let  $P_{\boldsymbol{\theta}_0}[\lim_{T\to\infty} \Pi\{\boldsymbol{\theta}: D_{\ell}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) < \epsilon \mid \mathbf{y}_{1:T}\} = 1] = 1$  for  $\ell \geq (2C^q + 1)q$ . Then, for all neighborhood  $U_{\mathbf{p}_0}$  of  $\mathbf{p}_0$  and all weak neighborhood  $U_{f_{c,0}}$  of  $f_{c,0}$ ,  $c = 1, \dots, C$ ,

$$P_{\boldsymbol{\theta}_0}\left[\lim_{T\to\infty}\Pi\left\{\boldsymbol{\theta}=(\mathbf{p},\mathbf{f}):\exists\sigma,\sigma(\mathbf{p})\in U_{\mathbf{p}_0},\ f_{\sigma(c)}\in U_{f_{c,0}},c=1,\ldots,C\mid \mathbf{y}_{1:T}\right\}=1\right]=1,$$

where  $\sigma$  is a permutation with  $\sigma(\mathbf{p}) = \{p_{\sigma(c_{q+1})|\sigma(c_1),\dots,\sigma(c_q)}, c_j = 1,\dots,C, j = 1,\dots,q+1\}.$ 

Some details of the proofs, including coupling arguments, are provided below to show that the original calculations follow through in the HOHMM settings using the above first order representation with single step transitions. This also makes the article relatively self-contained.

## S.2.1 Background

The following results are directly used in the proofs. In what follows,  $\|\mu - \nu\|_{TV}$  denotes the total variation distance between two probability measures  $\mu$  and  $\nu$ . That is,  $\|\mu - \nu\|_{TV} = \sup_A |\mu(A) - \nu(A)| = (1/2) \|\mu - \nu\|_{L_1}$ . For a sequence  $\{X_n\}_{n\geq 0}$ , its  $\phi$  mixing coefficients are defined as  $\phi_m = \sup_{A,B:P(A)>0} \{P(X_m \in B) - P(X_m \in B \mid X_0 \in A)\}$ .

Theorem 4. (Coupling Inequality, Doeblin, 1938, Lindvall, 1992 and Thorisson, 2000) For any coupling  $\Lambda$  of random variables  $X \sim \mu$  and  $Y \sim \nu$  we have

$$\|\mu - \nu\|_{TV} \le P_{\Lambda}(X \ne Y).$$

Consider a finite state Markov chain with state space  $C = \{1, ..., C\}$  and transition probability matrix P. Let  $P(x,y) \geq \epsilon_0 \psi(y)$  for some probability measure  $\psi(\cdot)$ . We can 'split' the transition kernel P into two constituent parts, one of which does not depend on the current position in the chain, using the decomposition

$$P(x,\cdot) = \epsilon_0 \psi(\cdot) + (1 - \epsilon_0) R(x,\cdot),$$

where  $R(x,\cdot) = \{P(x,\cdot) - \epsilon_0 \psi(\cdot)\}/(1 - \epsilon_0)$  is also a transition kernel. Generating the next sample in a Markov chain with kernel P can therefore be done in two stages. First, draw a Bernoulli random variable with probability of success  $\epsilon_0$ , and then conditional on success draw from  $\psi(\cdot)$ , otherwise draw from  $R(x,\cdot)$ .

Consider the following coupling with the Markov chains  $\{X_t\}_{t\geq 0}$  and  $\{Y_t\}_{t\geq 0}$ , both with state space  $\mathcal{C} = \{1, \ldots, C\}$ , transition kernel P and invariant distribution  $\pi$ .

- 1. For the chain  $\{X_t\}_{t\geq 0}$ , set  $X_0=x_0$ , and for  $\{Y_t\}_{t\geq 0}$ , set  $Y_0\sim \pi$ .
- 2. At each iteration m, draw  $U_m \sim \text{Bernoulli}(\epsilon_0)$ .
- 3. If  $U_m = 0$ , then draw  $X_m \sim R(x_{m-1}, \cdot)$  and  $Y_m \sim R(y_{m-1}, \cdot)$  independently.
- 4. But if  $U_m = 1$ , draw  $X_m \sim \psi$ , and set  $Y_m = X_m$ . In subsequent steps, draw values of  $X_t$  as above but always set  $Y_t = X_t$ .

We then have

$$P(X_1 \neq Y_1 \mid X_0, Y_0) = P(X_1 \neq Y_1 \mid U_1 = 0, X_0, Y_0)(1 - \epsilon_0) + P(X_1 \neq Y_1 \mid U_1 = 1, X_0, Y_0)\epsilon_0$$
  
=  $P(X_1 \neq Y_1 \mid U_1 = 0, X_0, Y_0)(1 - \epsilon_0) \le (1 - \epsilon_0).$ 

Under the coupling, once  $X_s = Y_s$  we have  $X_t = Y_t$  for all t > s. Therefore,

$$P(X_m \neq Y_m \mid X_0, Y_0) = P(X_m \neq Y_m, X_{m-1} \neq Y_{m-1}, \dots, X_1 \neq Y_1 \mid X_0, Y_0)$$

$$= P(X_m \neq Y_m \mid X_{m-1} \neq Y_{m-1}) \cdots P(X_2 \neq Y_2 \mid X_1 \neq Y_1) P(X_1 \neq Y_1 \mid X_0, Y_0)$$
  
 
$$\leq (1 - \epsilon_0)^m.$$

The bound is uniform since it does not depend on the starting value  $X_0$ . Therefore, we have

$$\phi_m[\{X_t\}_{t\geq 0}] \leq \sup_{x} \|P^m(x,\cdot) - \pi(\cdot)\|_{TV} \leq P_{\Lambda}(X_m \neq Y_m) \leq (1 - \epsilon_0)^m.$$

If P be such that  $P(x,y) \ge \underline{p} > 0$  for all  $x,y \in \mathcal{C}$ . Choosing  $\psi(y) = 1/C$  for all y, we can then set  $\epsilon_0 = Cp$ . Therefore, in this case, we have

$$\phi_m[\{X_t\}_{t\geq 0}] \leq \sup_x \|P^m(x,\cdot) - \pi(\cdot)\|_{TV} \leq (1 - C\underline{p})^m.$$

Consider now a stationary HMM with latent state sequence  $X_t$  and observation sequence  $Y_t$ . Then, for the Markov sequence  $\{(X_t, Y_t)\}_{t>0}$ , we have

$$\begin{split} \phi_m[\{(X_t,Y_t)\}_{t\geq 0}] &\leq \sup_{x_0,y_0,B=B_x\times B_y} |P\{(X_m,Y_m)\in B\} - P\{(X_m,Y_m)\in B\mid (X_0,Y_0)=(x_0,y_0)\}| \\ &= \sup_{x_0,y_0,B=B_x\times B_y} |\pi(X_m\in B_x)P(Y_m\in B_y\mid X_m\in B_x) - P(X_m\in B_x\mid X_0=x_0)P(Y_m\in B_y\mid X_m\in B_x)| \\ &= \sup_{x_0,y_0,B=B_x\times B_y} |P(Y_m\in B_y\mid X_m\in B_x)\{\pi(X_m\in B_x) - P(X_m\in B_x\mid X_0=x_0)\}| \\ &\leq \sup_{x_0,B_x} |\pi(B_x) - P^m(B_x\mid X_0=x_0)| \leq (1-C\underline{p})^m. \end{split}$$

The mixing coefficients  $\phi_m[\{(X_t, Y_t)\}_{t\geq 0}]$  of the Markov sequence  $\{(X_t, Y_t)\}_{t\geq 0}$  thus admit the same bounds as those on the mixing coefficients of the original sequence  $\{X_t\}_{t\geq 0}$ .

Consider next the Markov sequence  $\{Z_t\}_{t\geq 0}$  formed with blocks of the sequence  $\{(X_t,Y_t)\}_{t\geq 0}$  of size  $\ell$ , that is,  $Z_t = (X_{t\ell+1},\ldots,X_{t\ell+\ell},Y_{t\ell+1},\ldots,Y_{t\ell+\ell})$ , where  $\{X_t\}_{t\geq 0}$  and  $\{Y_t\}_{t\geq 0}$  are the latent and observed processes of an HMM as above. Then, proceeding as above, for the sequence  $\{Z_t\}_{t\geq 0}$ , we have

$$\phi_{m}[\{Z_{t}\}_{t\geq0}] \leq \sup_{x_{1},\dots,x_{\ell},B_{x,1},\dots,B_{x,\ell}} |P(X_{m\ell+1} \in B_{x,1},\dots,X_{m\ell+\ell} \in B_{x,\ell})$$

$$-P(X_{m\ell+1} \in B_{x,1},\dots,X_{m\ell+\ell} \in B_{x,\ell} \mid X_{1} = x_{1},\dots,X_{\ell} = x_{\ell})|$$

$$\leq \sup_{x_{\ell},B_{x,1},B_{x,2},\dots,B_{x,\ell}} |P(X_{m\ell+2} \in B_{x,2},\dots,X_{m\ell+\ell} \in B_{x,\ell} \mid X_{m\ell+1} \in B_{x,1})$$

$$\{P(X_{m\ell+1} \in B_{x,1}) - P(X_{m\ell+1} \in B_{x,1} \mid X_{\ell} = x_{\ell})\}|$$

$$\leq \sup_{x_{\ell},B_{x,1}} |P(X_{m\ell+1} \in B_{x,1}) - P(X_{m\ell+1} \in B_{x,1} \mid X_{\ell} = x_{\ell})|$$

$$\leq (1 - C\underline{p})^{(m-1)\ell+1} \leq (1 - C\underline{p})^{m}.$$

Next, consider a  $q^{th}$  order Markov chain  $\{X_t\}_{t\geq 0}$  with transition probabilities  $p(x_{q+1} \mid x_1, \ldots, x_q)$  with  $p(x_{q+1} \mid x_1, \ldots, x_q) \geq \underline{p}$  for all  $x_1, \ldots, x_q, x_{q+1} \in \mathcal{C}$ . We can still 'split' the transition kernel p into two constituent parts using the decomposition

$$p\{(x_1,\ldots,x_q)\to\cdot\}=\epsilon_0\psi(\cdot)+(1-\epsilon_0)R\{(x_1,\ldots,x_q)\to\cdot\}.$$

We now consider the following coupling for two higher order Markov chains  $\{X_t\}_{t\geq 0}$  and  $\{Y_t\}_{t\geq 0}$ , both with transition kernel p, stationary distribution  $\pi$ , marginal one component stationary distribution  $\pi_1$ , and special initial states as described below.

- 1. For  $\{X_t\}_{t\geq 0}$ , set  $X_0=(x_{-q+1},\ldots,x_0)$ , and for  $\{Y_t\}_{t\geq 0}$ , set  $Y_0=(y_{-q+1},\ldots,y_0)\sim\pi$ .
- 2. At each iteration m, draw  $U_m \sim \text{Bernoulli}(\epsilon_0)$ .
- 3. If  $U_m = 0$ , draw  $X_m \sim R\{(x_{m-q}, \dots, x_{m-1}) \to \cdot\}$  and  $Y_m \sim R\{(y_{m-q}, \dots, y_{m-1}) \to \cdot\}$  independently.
- 4. But if  $U_m = 1$ , draw  $X_m \sim \psi(\cdot)$ , and set  $Y_m = X_m$ . In subsequent steps, draw values of  $X_t$  as before but always set  $Y_t = X_t$ .

Following an argument and calculations similar to the first order case, we then have

$$\phi_m[\{X_t\}_{t\geq 0}] \leq \sup_{x_1,\dots,x_q} \|P^m\{(x_1,\dots,x_q),\cdot\} - \pi_1(\cdot)\|_{TV} \leq (1-\epsilon_0)^m = (1-C\underline{p})^m.$$

The sequence  $\{Z_j\}_{j\geq 0}$ , where  $Z_j=(X_{j\ell+1},...,X_{j\ell+\ell},Y_{j\ell+1},...,Y_{j\ell+\ell})$  is also a Markov sequence whenever  $\ell\geq q$ . Proceeding as above, under the assumption that  $p(x_{q+1}\mid x_1,...,x_q)\geq p$  for all  $x_1,...,x_q,x_{q+1}\in\mathcal{C}$ , we have

$$\phi_{m}[\{Z_{t}\}_{t\geq 1}] \leq \sup_{x_{1},\dots,x_{\ell},B_{x,1},\dots,B_{x,\ell}} |P(X_{m\ell+1} \in B_{x,1},\dots,X_{m\ell+\ell} \in B_{x,\ell})$$

$$-P(X_{m\ell+1} \in B_{x,1},\dots,X_{m\ell+\ell} \in B_{x,\ell} \mid X_{1} = x_{1},\dots,X_{\ell} = x_{\ell})|$$

$$\leq \sup_{x_{\ell-q},\dots,x_{\ell},B_{x,1},B_{x,2},\dots,B_{x,\ell}} |P(X_{m\ell+2} \in B_{x,2},\dots,X_{m\ell+\ell} \in B_{x,\ell} \mid X_{m\ell+1} \in B_{x,1})$$

$$\{P(X_{m\ell+1} \in B_{x,1}) - P(X_{m\ell+1} \in B_{x,1} \mid X_{\ell-q} = x_{\ell-q},\dots,X_{\ell} = x_{\ell})\}|$$

$$\leq (1 - C\underline{p})^{(m-1)\ell+1} \leq (1 - C\underline{p})^{m}.$$

## Theorem 5 of Barron, 1988)

a. Let  $p_{T, \boldsymbol{\theta}_0}(\mathbf{y}_{1:T})$  and  $\int_{\Theta} p_{T, \boldsymbol{\theta}}(\mathbf{y}_{1:T}) \Pi(d\boldsymbol{\theta})$  merge with probability one. That is, for every  $\beta > 0$ 

$$P_{\pmb{\theta}_0} \left\{ \frac{\int_{\Theta} p_{T, \pmb{\theta}}(\mathbf{y}_{1:T}) \Pi(d\pmb{\theta})}{p_{T, \pmb{\theta}_0}(\mathbf{y}_{1:T})} \leq \exp(-T\beta) \ \textit{infintely often} \right\} = 0.$$

- b. The sets  $A_T$ ,  $B_T$ ,  $C_T$  are such that
  - b.1.  $A_T \cup B_T \cup C_T = \Theta$ ,
  - b.2.  $\Pi(B_T) \leq \exp(-T\beta_1)$  for some  $\beta_1 > 0$ ,
  - b.3. there exists a uniformly exponentially consistent test  $\psi_T(\mathbf{y}_1, \dots, \mathbf{y}_T)$  for testing  $H_0: f = f(\boldsymbol{\theta}_0)$  vs  $H_1: f \in \{f(\boldsymbol{\theta}): \boldsymbol{\theta} \in C_T\}$ . That is, for some  $\beta_2 > 0$ ,

$$\lim_{T \to \infty} \mathbb{E}_{\boldsymbol{\theta}_0}(\psi_T) = 0 \quad and \quad \sup_{\boldsymbol{\theta} \in C_T} \mathbb{E}_{\boldsymbol{\theta}}(1 - \psi_T) \le \exp(-T\beta_2).$$

Then,  $P_{\theta_0}\{\lim_{T\to\infty} \Pi(A_T \mid \mathbf{y}_{1:T}) = 1\} = 1.$ 

#### Theorem 6. (Rio's inequality, Rio, 2000)

Let  $\{X_n\}_{n\geq 0}$  be a sequence of random variables, each  $X_i$  taking values in  $E_i$ , with mixing coefficients  $\phi_m = \sup_{A,B:P(A)>0} \{P(X_m \in B) - P(X_m \in B \mid X_0 \in A)\}$ . Let f be a function from  $E^n = E_1 \times \cdots \times E_n$  to  $\mathbb{R}$  such that

$$|f(x_1,\ldots,x_n)-f(y_1,\ldots,y_n)| \le \Delta_1 1_{x_1 \ne y_1} + \cdots + \Delta_n 1_{x_n \ne y_n}.$$

Then, for any x > 0

$$P\{f(X_1,\ldots,X_n) - \mathbb{E}_P f(X_1,\ldots,X_n) \ge x\} \le \exp\left\{-\frac{2x^2(\Delta_1^2 + \cdots + \Delta_n^2)^{-1}}{(1 + 2\phi_1 + \cdots + 2\phi_{n-1})^2}\right\}.$$

For a stationary Markov sequence with transition probabilities  $p(x_2 \mid x_1) \geq \underline{p}$  for all  $x_1, x_2 \in \mathcal{C}$ , we have  $\phi_m \leq \rho_{\boldsymbol{\theta}}^{-m}$  with  $\rho_{\boldsymbol{\theta}} \geq (1 - C\underline{p})^{-1}$ . Therefore,  $(1 + 2\phi_1 + \cdots + 2\phi_{n-1}) \leq (1 + 2\rho_{\boldsymbol{\theta}}^{-1} + 2\rho_{\boldsymbol{\theta}}^{-2} + \cdots) = (2 + \rho_{\boldsymbol{\theta}} - 1)/(\rho_{\boldsymbol{\theta}} - 1) \leq (2 - C\underline{p})/(C\underline{p})$ . Letting  $f(x_1, \dots, x_n) = \sum_{i=1}^n h(x_i)$  with  $0 \leq h(x) \leq 1$  so that  $\Delta_i = 1$  for all i, we then have

$$P\left[\sum_{i=1}^{n} \left\{h(X_i) - \mathbb{E}_P h(X_i)\right\} \ge x\right] \le \exp\left\{-\frac{2x^2 C^2 \underline{p}^2}{n(2 - C\underline{p})^2}\right\} \le \exp\left\{-\frac{x^2 C^2 \underline{p}^2}{2n}\right\}.$$

Following the discussions above, the same inequality then also holds true for the sequence  $\{Z_t\}_{t\geq 0}$  with  $Z_t = (X_{t\ell+1}, \ldots, X_{t\ell+\ell}, Y_{t\ell+1}, \ldots, Y_{t\ell+\ell})$ , where  $\{X_t\}_{t\geq 0}$  and  $\{Y_t\}_{t\geq 0}$  are the latent and observed processes of an HOHMM with transition kernel of the latent sequence  $\{X_t\}_{t\geq 0}$  satisfying  $p(x_{q+1} \mid x_1, \ldots, x_q) \geq p$  for all  $x_1, \ldots, x_q, x_{q+1} \in \{1, \ldots, C\}$ .

#### S.2.2 Proof of Theorem 2

The likelihood function of an HOHMM is given by

$$p_{T,\boldsymbol{\theta}} = p(\mathbf{y}_{1:T} \mid \boldsymbol{\theta}) = \sum_{c_1, \dots, c_T} \pi(c_1, \dots, c_q) \prod_{t=q+1}^T p(c_t \mid \mathbf{c}_{(t-q):(t-1)}) \prod_{t=1}^T f(\mathbf{y}_t \mid c_t),$$

where  $\pi_{c_1,\dots,c_q}$  is the stationary as well as the starting distribution of the HOHMM. Following the steps similar to those the proof of Lemma 2.2 in Vernet (2015b), we can show, under Assumptions 1, that the Kullback-Leibler divergence between  $p_{T,\boldsymbol{\theta}_0}$  and  $p_{T,\boldsymbol{\theta}}$ , denoted  $KL(p_{T,\boldsymbol{\theta}_0}, p_{T,\boldsymbol{\theta}})$ , satisfies

$$KL\left(p_{T,\boldsymbol{\theta}_{0}}, p_{T,\boldsymbol{\theta}}\right) = \mathbb{E}_{\boldsymbol{\theta}_{0}}\log\left\{\frac{\sum_{c_{1},...,c_{T}}\pi_{c_{1},...,c_{q},0}\prod_{t=q+1}^{T}p_{c_{t}|\mathbf{c}_{(t-q):(t-1),0}}\prod_{t=1}^{T}f_{c_{t},0}(\mathbf{y}_{t})}{\sum_{c_{1},...,c_{T}}\pi_{c_{1},...,c_{q}}\prod_{t=q+1}^{T}p_{c_{t}|\mathbf{c}_{(t-q):(t-1)}}\prod_{t=1}^{T}f_{c_{t}}(\mathbf{y}_{t})}\right\}$$

$$\leq \max_{\mathbf{c}_{1:q}}\left|\pi_{c_{1},...,c_{q},0}-\pi_{c_{1},...,c_{q}}\right|/\underline{p}+(T-q)\max_{\mathbf{c}_{1:(q+1)}}\left|p_{c_{q+1}|c_{1},...,c_{q},0}-p_{c_{q+1}|c_{1},...,c_{q}}\right|$$

$$+T\max_{c}\int f_{c,0}(\mathbf{y})\max_{k}\log\frac{f_{c,0}(\mathbf{y})}{f_{k}(\mathbf{y})}\lambda(\mathbf{y}).$$

This implies, for any  $\epsilon > 0$  and any  $\theta \in \Theta_{\epsilon}$ , that

$$\frac{1}{T}KL\left(p_{T,\pmb{\theta}_0},p_{T,\pmb{\theta}}\right) \leq \frac{3}{p}\epsilon.$$

Using the bound on the KL divergence, following the steps in the proof of Theorem 2.1 in Vernet (2015b), we can show that for any  $\beta > 0$ 

$$P_{\boldsymbol{\theta}_0} \left\{ \frac{\int_{\Theta} p_{T,\boldsymbol{\theta}}(\mathbf{y}_{1:T}) \Pi(d\boldsymbol{\theta})}{p_{T,\boldsymbol{\theta}_0}(\mathbf{y}_{1:T})} < \exp(-\beta T) \text{ infinitely often} \right\} = 0.$$

Consistency on any weak neighborhood of  $f_{1:\ell,0}^{\star}$  is equivalent to consistency in neighborhoods of the type  $U_h = \{f_{1:\ell}^{\star}: \int h f_{1:\ell}^{\star} - \int h f_{1:\ell,0}^{\star} < \epsilon \}$  where  $0 \leq h \leq 1$  is a bounded continuous function on  $\mathcal{Y}^{\ell}$  (Ghosh and Ramamoorthi, 2003, page 131). Uniformly exponentially consistent tests for  $H_0: f_{1:\ell}^{\star} = f_{1:\ell,0}^{\star}$  vs  $H_1: f_{1:\ell}^{\star} \in U_h^c$  may then be constructed as  $\psi_T(\mathbf{y}_{1:T}) = 1_{\mathcal{S}_T}(\mathbf{y}_{1:T})$  where

$$\mathcal{S}_T = \left\{ \mathbf{y}_{1:T} : \frac{\ell}{T} \sum_{j=0}^{T/\ell-1} h(\mathbf{y}_{j\ell+1}, \dots, \mathbf{y}_{j\ell+\ell}) > \frac{\alpha + \gamma}{2} \right\},\,$$

$$\alpha = \mathbb{E}_{f_{1:\ell,0}^{\star}}\{h(\mathbf{y}_1,\ldots,\mathbf{y}_{\ell})\}, \qquad \gamma = \inf_{f_{1:\ell}^{\star} \in U_b^{\star}} \mathbb{E}_{f_{1:\ell}^{\star}}\{h(\mathbf{y}_1,\ldots,\mathbf{y}_{\ell})\} > \alpha + \epsilon.$$

Then, using Rio's inequality, we have

$$\mathbb{E}_{\boldsymbol{\theta}_{0}}(\psi_{T}) = P_{\boldsymbol{\theta}_{0}}(\mathcal{S}_{T}) 
= P_{\boldsymbol{\theta}_{0}} \left[ \sum_{j=0}^{T/\ell-1} \left\{ h(\mathbf{y}_{j\ell+1}, \dots, \mathbf{y}_{j\ell+\ell}) - \int h(\mathbf{y}_{j\ell+1}, \dots, \mathbf{y}_{j\ell+\ell}) f_{1:\ell,0}^{\star}(\mathbf{y}_{j\ell+1}, \dots, \mathbf{y}_{j\ell+\ell}) d\lambda^{\otimes \ell} \right\} > \frac{T(\gamma - \alpha)}{2\ell} \right] 
\leq \exp \left\{ -\frac{T(\gamma - \alpha)^{2} C^{2} \underline{p}^{2}}{32\ell} \right\}.$$

Likewise, for all  $\boldsymbol{\theta} \in \Theta(p)$  such that  $f_{1:\ell}^{\star} \in U_h^c$ 

$$\mathbb{E}_{\boldsymbol{\theta}}(1 - \psi_{T}) = P_{\boldsymbol{\theta}}(\mathcal{S}_{T}^{c}) 
= P_{\boldsymbol{\theta}} \left[ \sum_{j=0}^{T/\ell-1} \left\{ -h(\mathbf{y}_{(j\ell+1):(j\ell+\ell)}) + \mathbb{E}_{f_{1:\ell}^{\star}} h(\mathbf{y}_{(j\ell+1):(j\ell+\ell)}) \right\} > \frac{T}{\ell} \mathbb{E}_{f_{1:\ell}^{\star}} h(\mathbf{y}_{(j\ell+1):(j\ell+\ell)}) - \frac{T(\gamma + \alpha)}{2\ell} \right] 
\leq P_{\boldsymbol{\theta}} \left[ \sum_{j=0}^{T/\ell-1} \left\{ -h(\mathbf{y}_{j\ell+1}, \dots, \mathbf{y}_{j\ell+\ell}) + \int h(\mathbf{y}_{j\ell+1}, \dots, \mathbf{y}_{j\ell+\ell}) f_{1:\ell}^{\star}(\mathbf{y}_{j\ell+1}, \dots, \mathbf{y}_{j\ell+\ell}) d\lambda^{\otimes \ell} \right\} > \frac{T(\gamma - \alpha)}{2\ell} \right] 
\leq \exp \left\{ -\frac{T(\gamma - \alpha)^{2} C^{2} \underline{p}^{2}}{32\ell} \right\}.$$

Applying Barron's theorem with  $A_T = U_h$ ,  $B_T = \phi$ ,  $C_T = U_h^c$ , we have that  $P_{\theta_0}\{\lim_{T\to\infty} \Pi(U_h \mid \mathbf{y}_{1:T}) = 1\} = 1$ . This proves Theorem 2.

#### S.2.3 Proof of Theorem 1

Let  $\|f_{1:\ell}^{\star} - f_{1:\ell,0}^{\star}\|_1 = \int |f_{1:\ell}^{\star} - f_{1:\ell,0}^{\star}| \lambda^{\otimes \ell}$  and  $\|\mathbf{p}_0 - \mathbf{p}\| = \max_{\mathbf{c}_{(t-q):t}} |p_{c_t|c_{t-q},\dots,c_{t-1},0} - p_{c_t|c_{t-q},\dots,c_{t-1}}|$ . We now let  $U = \{\boldsymbol{\theta} : D_{\ell}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) < \epsilon\}$  denote a  $D_{\ell}$  neighborhood of  $\boldsymbol{\theta}_0$ . Since  $\Pi\{\Theta(\underline{p})\} = 1$ , in what follows, complements are implicitly understood to be with respect the set  $\Theta(\underline{p})$ . That is,  $U^c = \Theta(p) \cap U^c$  and so on. To apply Barron's theorem, we then set

$$A_T = U$$
,  $B_T = \mathcal{P}(p) \times \mathcal{F}_T$ ,  $C_T = (A_T \cup B_T)^c$ .

Using Assumption 1C from the main paper, we have that

$$\Pi(B_T) = \Pi_F(\mathcal{F}_T^c) \le \exp(-T\beta_1).$$

We next have to show the existence of a uniformly consistent test  $\psi_T$  for testing  $H_0: f_{1:\ell}^{\star} = f_{1:\ell,0}^{\star}$  vs  $H_1: f_{1:\ell}^{\star} \in C_T$ .

Let  $f_{1:\ell,j}^{\star} = f_{1:\ell}^{\star}(\mathbf{y}_1, \dots, \mathbf{y}_{\ell} \mid \boldsymbol{\theta}_j)$ . Let  $\boldsymbol{\theta}_j, j = 1, \dots, N, \ N = \mathcal{N}(\delta, \mathcal{F}_T, D_{\ell})$  be a sequence such that for all  $\boldsymbol{\theta} \in \mathcal{P}(\underline{p}) \times \mathcal{F}_T, \ D_{\ell}(\boldsymbol{\theta}_j, \boldsymbol{\theta}) \leq \delta$  with  $\delta = \epsilon/12$ . Define  $\phi_j(\mathbf{y}_{1:T}) = 1_{\mathcal{S}_j}(\mathbf{y}_{1:T})$ , where

$$S_{j} = \left\{ \mathbf{y}_{1:T} : \sum_{j=0}^{T/\ell-1} \left[ 1\{ (\mathbf{y}_{j\ell+1}, \dots, \mathbf{y}_{j\ell+\ell}) \in B_{j} \} - P_{\boldsymbol{\theta}_{0}} \{ (\mathbf{y}_{1}, \dots, \mathbf{y}_{\ell}) \in B_{j} \} \right] > s_{j} \right\},$$
with  $B_{j} = \{ (\mathbf{y}_{1}, \dots, \mathbf{y}_{\ell}) \in \mathcal{Y}^{\ell} : f_{1:\ell}^{\star}(\mathbf{y}_{1}, \dots, \mathbf{y}_{\ell} \mid \boldsymbol{\theta}_{0}) \leq f_{1:\ell}^{\star}(\mathbf{y}_{1}, \dots, \mathbf{y}_{\ell} \mid \boldsymbol{\theta}_{j}) \}$ 
and  $s_{j} = \frac{T \left\| f_{1:\ell,j}^{\star} - f_{1:\ell,0}^{\star} \right\|_{1}}{4\ell}.$ 

Then, we have

$$P_{\boldsymbol{\theta}_{j}}\{(\mathbf{y}_{j\ell+1},\ldots,\mathbf{y}_{j\ell+\ell})\in B_{j}\}-P_{\boldsymbol{\theta}_{0}}\{(\mathbf{y}_{1},\ldots,\mathbf{y}_{\ell})\in B_{j}\}=\frac{1}{2}\left\|f_{1:\ell,j}^{\star}-f_{1:\ell,0}^{\star}\right\|_{1}.$$

Using Rio's inequality again, we have

$$\mathbb{E}_{\boldsymbol{\theta}_{0}}(\phi_{j}) = P_{\boldsymbol{\theta}_{0}}(\mathcal{S}_{j}) = P_{\boldsymbol{\theta}_{0}} \left[ \sum_{j=0}^{T/\ell-1} \left\{ 1(\mathbf{y}_{(j\ell+1):(j\ell+\ell)} \in B_{j}) - P_{\boldsymbol{\theta}_{0}}(\mathbf{y}_{1:\ell} \in B_{j}) \right\} > s_{j} \right]$$

$$\leq \exp \left\{ -\frac{T \left\| f_{1:\ell,j}^{\star} - f_{1:\ell,0}^{\star} \right\|_{1}^{2} C^{2} \underline{p}^{2}}{32\ell} \right\}.$$

We next define

$$\psi_T = \max_{1 \le j \le N: \boldsymbol{\theta}_j \in A_T^c} \phi_j.$$

For  $\boldsymbol{\theta}_j \in A_T^c$ ,  $\left\| f_{1:\ell,j}^{\star} - f_{1:\ell,0}^{\star} \right\|_1 \ge \epsilon$ . We then have

$$\mathbb{E}_{\boldsymbol{\theta}_{0}}(\psi_{T}) = \mathbb{E}_{\boldsymbol{\theta}_{0}}\left(\max_{1 \leq j \leq N: \boldsymbol{\theta}_{j} \in A_{T}^{c}} \phi_{j}\right) \leq \mathbb{E}_{\boldsymbol{\theta}_{0}}\left(\sum_{1 \leq j \leq N: \boldsymbol{\theta}_{j} \in A_{T}^{c}} \phi_{j}\right) = \sum_{1 \leq j \leq N: \boldsymbol{\theta}_{j} \in A_{T}^{c}} \mathbb{E}_{\boldsymbol{\theta}_{0}}\left(\phi_{j}\right)$$

$$\leq \mathcal{N}\left\{\frac{\epsilon}{12}, \mathcal{P}(\underline{p}) \times \mathcal{F}_{T}, D_{\ell}\right\} \max_{1 \leq j \leq N: \boldsymbol{\theta}_{j} \in A_{T}^{c}} \mathbb{E}_{\boldsymbol{\theta}_{0}}(\phi_{j})$$

$$\leq \mathcal{N}\left\{\frac{\epsilon}{12}, \mathcal{P}(\underline{p}) \times \mathcal{F}_{T}, D_{\ell}\right\} \exp\left(-\frac{T\epsilon^{2}C^{2}\underline{p}^{2}}{32\ell}\right).$$

Likewise, we have

$$\mathbb{E}_{\boldsymbol{\theta}}(1 - \phi_{j}) = P_{\boldsymbol{\theta}}(\mathcal{S}_{j}^{c}) 
= P_{\boldsymbol{\theta}} \left[ \sum_{j=0}^{T/\ell-1} \left\{ -1(\mathbf{y}_{(j\ell+1):(j\ell+\ell)} \in B_{j}) + P_{\boldsymbol{\theta}_{0}}(\mathbf{y}_{1:\ell} \in B_{j}) \right\} > -s_{j} \right] 
= P_{\boldsymbol{\theta}} \left[ \sum_{j=0}^{T/\ell-1} \left\{ -1(\mathbf{y}_{(j\ell+1):(j\ell+\ell)} \in B_{j}) + P_{\boldsymbol{\theta}}(\mathbf{y}_{1:\ell} \in B_{j}) \right\} \right] 
> -s_{j} + \sum_{j=0}^{T/\ell-1} \left\{ P_{\boldsymbol{\theta}}(\mathbf{y}_{(j\ell+1):(j\ell+\ell)} \in B_{j}) - P_{\boldsymbol{\theta}_{0}}(\mathbf{y}_{1:\ell} \in B_{j}) \right\} \right].$$

Now, for all  $\boldsymbol{\theta} \in A_T^c$ , we have

$$-s_{j} + \sum_{j=0}^{T/\ell-1} \left\{ P_{\boldsymbol{\theta}}(\mathbf{y}_{(j\ell+1):(j\ell+\ell)} \in B_{j}) - P_{\boldsymbol{\theta}_{0}}(\mathbf{y}_{1:\ell} \in B_{j}) \right\}$$

$$= -\frac{T \left\| f_{1:\ell,j}^{\star} - f_{1:\ell,0}^{\star} \right\|_{1}}{4\ell} + \sum_{j=0}^{T/\ell-1} \left\{ P_{\boldsymbol{\theta}_{j}}(\mathbf{y}_{(j\ell+1):(j\ell+\ell)} \in B_{j}) - P_{\boldsymbol{\theta}_{0}}(\mathbf{y}_{1:\ell} \in B_{j}) \right\}$$

$$+ \sum_{j=0}^{T/\ell-1} \left\{ P_{\boldsymbol{\theta}}(\mathbf{y}_{(j\ell+1):(j\ell+\ell)} \in B_{j}) - P_{\boldsymbol{\theta}_{j}}(\mathbf{y}_{1:\ell} \in B_{j}) \right\}$$

$$= -\frac{T \left\| f_{1:\ell,j}^{\star} - f_{1:\ell,0}^{\star} \right\|_{1}}{4\ell} + \frac{T \left\| f_{1:\ell,j}^{\star} - f_{1:\ell,0}^{\star} \right\|_{1}}{2\ell} - \frac{T \left\| f_{1:\ell,j}^{\star} - f_{1:\ell}^{\star} \right\|_{1}}{2\ell}$$

$$= \frac{T \left\| f_{1:\ell,j}^{\star} - f_{1:\ell,0}^{\star} \right\|_{1}}{4\ell} - \frac{T \left\| f_{1:\ell,j}^{\star} - f_{1:\ell}^{\star} \right\|_{1}}{2\ell} \ge \frac{T\epsilon}{4\ell} - \frac{T\epsilon}{24\ell} \ge \frac{T\epsilon}{8\ell}.$$

The last line follows since for all  $\boldsymbol{\theta} \in A_T^c$ , we have  $\|f_{1:\ell}^{\star} - f_{1:\ell,0}^{\star}\|_1 > \epsilon$ , and also by definition of  $\boldsymbol{\theta}_j$ , we have  $\|f_{1:\ell,j}^{\star} - f_{1:\ell}^{\star}\|_1 < \epsilon/12$ . Therefore, for any  $\boldsymbol{\theta} \in A_T^c$ , using Rio's inequality, we have

$$\mathbb{E}_{\boldsymbol{\theta}}(1-\phi_j) \leq P_{\boldsymbol{\theta}}\left[\sum_{j=0}^{T/\ell-1} \left\{-1(\mathbf{y}_{(j\ell+1):(j\ell+\ell)} \in B_j) + P_{\boldsymbol{\theta}}(\mathbf{y}_{1:\ell} \in B_j)\right\} > \frac{T\epsilon}{8\ell}\right] \leq \exp\left(-\frac{T\epsilon^2 C^2 \underline{p}^2}{32\ell}\right).$$

Therefore, we have

$$\sup_{\boldsymbol{\theta} \in A_T^c \cap \mathcal{B}_T^c} \mathbb{E}_{\boldsymbol{\theta}} (1 - \psi_T) \le \exp\left(-\frac{T\epsilon^2 C^2 \underline{p}^2}{32\ell}\right).$$

For all  $\theta_0$ ,  $\theta$ , we have, using triangle inequality, that

$$D_{\ell}(\boldsymbol{\theta}_{0}, \boldsymbol{\theta}) \leq \sum_{c_{1}, \dots, c_{q}} \left| \pi_{c_{1}, \dots, c_{q}, 0} - \pi_{c_{1}, \dots, c_{q}} \right| + C(\ell - q) \|\mathbf{p}_{0} - \mathbf{p}\| + \ell \max_{c} \int |f_{c, 0}(\mathbf{y}_{1}) - f_{c}(\mathbf{y}_{1})| \lambda(d\mathbf{y}).$$

Therefore,  $D_{\ell}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \leq \epsilon/12$  if all three terms on the right hand side above are  $\leq \epsilon/36$ . The function  $\mathbf{p} \to \boldsymbol{\pi}(\mathbf{p})$  is uniformly continuous on  $\mathcal{P}(\underline{p})$ . Hence, there exists  $\epsilon_1 > 0$  such that for all  $\mathbf{p}_0, \mathbf{p} \in \mathcal{P}(\underline{p})$  with  $\|\mathbf{p}_0 - \mathbf{p}\| < \epsilon_1$ , we have  $\sum_{c_1, \dots, c_q} |\pi_{c_1, \dots, c_q, 0} - \pi_{c_1, \dots, c_q}| < \epsilon/36$ . This implies

$$\mathcal{N}\left\{\frac{\epsilon}{12}, \mathcal{P}(\underline{p}) \times \mathcal{F}_{T}, D_{\ell}\right\} \leq \mathcal{N}\left[\min\left\{\frac{\epsilon}{36C(\ell-q)}, \epsilon_{1}\right\}, \Theta(\underline{p}), \|\cdot\|\right] \mathcal{N}\left\{\frac{\epsilon}{36\ell}, \mathcal{F}_{T}, d(\cdot, \cdot)\right\} \\
\leq \left[\max\left\{\frac{36C(\ell-q)}{\epsilon}, \frac{1}{\epsilon_{1}}\right\}\right]^{C^{q}(C^{q}-1)} \mathcal{N}\left\{\frac{\epsilon}{36\ell}, \mathcal{F}_{T}, d(\cdot, \cdot)\right\}.$$

Under Assumptions 1, this implies

$$\mathbb{E}_{\boldsymbol{\theta}_0}(\psi_T) \leq \mathcal{N}\left\{\frac{\epsilon}{12}, \mathcal{P}(\underline{p}) \times \mathcal{F}_T, D_{\ell}\right\} \exp\left(-\frac{T\epsilon^2 C^2 \underline{p}^2}{32\ell}\right) \to 0.$$

Applying Barron's theorem, we have that  $P_{\theta_0}\{\lim_{T\to\infty}\Pi(U\mid \mathbf{y}_{1:T})=1\}=1$ . This concludes the proof of Theorem 1.

#### S.2.4 Proof of Theorem 3

It suffices to show that for any  $U_{\mathbf{p}_0}$  and  $U_{f_{c,0}}$ , there exists a  $D_{\ell}$ -neighborhood U of  $\boldsymbol{\theta}_0$  such that

$$U \subset \{\boldsymbol{\theta} = (\mathbf{p}, \mathbf{f}) : \exists \sigma, \sigma(\mathbf{p}) \in U_{\mathbf{p}_0}, \ f_{\sigma(c)} \in U_{f_{c,0}}, c = 1, \dots, C \mid \mathbf{y}_{1:T} \}.$$

This is equivalent to showing that for any  $\boldsymbol{\theta}_T = (\mathbf{p}_T, \mathbf{f}_T) \in \Theta(\underline{p})$  with  $D_{T_0}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_T) \to 0$ ,  $T_0 = (2C^q + 1)q$ , there exists a subsequence  $\tilde{\boldsymbol{\theta}}_T = (\tilde{\mathbf{p}}_T, \tilde{\mathbf{f}}_T)$  of  $\boldsymbol{\theta}_T$  such that  $\|\tilde{\mathbf{p}}_T - \bar{\mathbf{p}}\| \to 0$  and  $\mathbf{f}_{c,T} \to \bar{f}_{c,0}$  in the weak topology for some  $\bar{\boldsymbol{\theta}} \in \Theta(\underline{p})$  where  $(\bar{\mathbf{p}}, \bar{\mathbf{f}})$  and  $(\mathbf{p}_0, \mathbf{f}_0)$  (and hence  $\bar{\boldsymbol{\pi}}$  and  $\boldsymbol{\pi}_0$ ) are equivalent up to label swapping of the states.

Let  $\boldsymbol{\theta}_T = (\mathbf{p}_T, \mathbf{f}_T) \in \Theta(\underline{p})$  be such that  $D_{T_0}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_T) \to 0$ . Since  $\mathcal{P}(\underline{p})$  is compact, there exists a subsequence of  $\mathbf{p}_T$ , denoted  $\tilde{\mathbf{p}}_T$ , such that  $\tilde{\mathbf{p}}_T \to \bar{\mathbf{p}} \in \mathcal{P}(\underline{p})$ . Let  $\tilde{\boldsymbol{\pi}}_T = \tilde{\boldsymbol{\pi}}_T(\tilde{\mathbf{p}}_T)$  denote the associated stationary distribution and  $\tilde{\mathbf{f}}_T$  the associated emission distributions. Using triangle inequality, we have

$$D_{T_0}(\boldsymbol{\theta}_0, \widetilde{\boldsymbol{\theta}}_T) = \int \left| \sum_{c_1, \dots, c_{T_0}} \widetilde{\pi}_{c_1, \dots, c_q, T} \prod_{t=q+1}^{T_0} \widetilde{p}_{c_t | c_{t-q}, \dots, c_{t-1}, T} \prod_{t=1}^{T_0} \widetilde{f}_{c_t, T}(\mathbf{y}_t) \right|$$

$$-\sum_{c_{1},\dots,c_{T_{0}}} \pi_{c_{1},\dots,c_{q},0} \prod_{t=q+1}^{T_{0}} p_{c_{t}|c_{t-q},\dots,c_{t-1},0} \prod_{t=1}^{T_{0}} f_{c_{t},0}(\mathbf{y}_{t}) \left| \lambda(d\mathbf{y}_{1}) \dots \lambda(d\mathbf{y}_{T_{0}}) \right|$$

$$\geq -\sum_{c_{1},\dots,c_{T_{0}}} \left| \pi_{c_{1},\dots,c_{q}} \prod_{t=q+1}^{T_{0}} p_{c_{t}|c_{t-q},\dots,c_{t-1}} - \bar{\pi}_{c_{1},\dots,c_{q}} \prod_{t=q+1}^{T_{0}} \bar{p}_{c_{t}|c_{t-q},\dots,c_{t-1}} \right|$$

$$+ \int \left| \sum_{c_{1},\dots,c_{T_{0}}} \bar{\pi}_{c_{1},\dots,c_{q}} \prod_{t=q+1}^{T_{0}} \bar{p}_{c_{t}|c_{t-q},\dots,c_{t-1}} \prod_{t=1}^{T_{0}} \widetilde{f}_{c_{t},T}(\mathbf{y}_{t}) \right|$$

$$- \sum_{c_{1},\dots,c_{T_{0}}} \pi_{c_{1},\dots,c_{q},0} \prod_{t=q+1}^{T_{0}} p_{c_{t}|c_{t-q},\dots,c_{t-1},0} \prod_{t=1}^{T_{0}} f_{c_{t},0}(\mathbf{y}_{t}) \left| \lambda(d\mathbf{y}_{1}) \dots \lambda(d\mathbf{y}_{T_{0}}) \right|$$

Since the left hand side and the first term on the right hand side both tend to zero, we have

$$\int \left| \sum_{c_1,\dots,c_{T_0}} \bar{\pi}_{c_1,\dots,c_q} \prod_{t=q+1}^{T_0} \bar{p}_{c_t|c_{t-q},\dots,c_{t-1}} \prod_{t=1}^{T_0} \widetilde{f}_{c_t,T}(\mathbf{y}_t) - \sum_{c_1,\dots,c_{T_0}} \pi_{c_1,\dots,c_q,0} \prod_{t=q+1}^{T_0} p_{c_t|c_{t-q},\dots,c_{t-1},0} \prod_{t=1}^{T_0} f_{c_t,0}(\mathbf{y}_t) \right| \lambda(d\mathbf{y}_1) \dots \lambda(d\mathbf{y}_{T_0}) \to 0.$$

Since  $\sum_{c_1,\dots,c_{T_0}} \bar{\pi}_{c_1,\dots,c_q} \prod_{t=q+1}^{T_0} \bar{p}_{c_t|c_{t-q},\dots,c_{t-1}} \prod_{t=1}^{T_0} \widetilde{f}_{c_t,T}(\mathbf{y}_t)$  converges in total variation, it is tight, and hence so are the sequence of distributions  $\widetilde{f}_{c,T}$ . Using Prokhorov's theorem (Ghosh and Ramamoorthi, 2003, page 13), there exist subsequences of  $\widetilde{f}_{c,T}$  that converge weakly to some  $\bar{f}_{c_t}$ . This implies

$$\sum_{c_1,\dots,c_{T_0}} \bar{\pi}_{c_1,\dots,c_q} \prod_{t=q+1}^{T_0} \bar{p}_{c_t|c_{t-q},\dots,c_{t-1}} \prod_{t=1}^{T_0} \bar{f}_{c_t,T}(\mathbf{y}_t) = \sum_{c_1,\dots,c_{T_0}} \pi_{c_1,\dots,c_q,0} \prod_{t=q+1}^{T_0} p_{c_t|c_{t-q},\dots,c_{t-1},0} \prod_{t=1}^{T_0} f_{c_t,0}(\mathbf{y}_t).$$

Application of Lemma 1 then concludes the proof of Theorem 3.

As we discussed in Section 2.4 in the main paper, for the proposed tensor decomposition based model (5) for HOHMM transition probabilities, truncated Dirichlet priors on the parameters  $\lambda_{h_1,\dots,h_q}$ , truncated below  $\underline{p}$ , satisfy the assumptions on the transition probabilities required in Theorem 1. Results showing how the additional assumptions on the emission distributions and associated priors in Theorem 1 relate to the specific examples considered in the main paper can be derived along the lines of similar results in Vernet (2015b) and are omitted.

# S.3 Higher Order Chinese Restaurant Franchise (CRF)

### S.3.1 The Original CRF

We first review the original CRF (Teh et al. 2006) before we describe how we adapted it to our HOHMM setting in the next subsection. Let there be J groups, each with  $N_j$  observations  $\{y_{j,\ell}\}_{\ell=1}^{N_j}$  with a generative model as

$$\lambda_0 \mid \alpha_0 \sim \text{Dir}(\alpha_0/C, \dots, \alpha_0/C), 
\lambda_j \mid \alpha, \lambda_0 \sim \text{Dir}(\alpha \lambda_0), \quad c_{j,\ell} \mid \lambda_j \sim \text{Mult}\{\lambda_j(1), \dots, \lambda_j(C)\} 
y_{j,\ell} \mid \{\theta_c\}_{c=1}^C, c_{j,\ell} = k \sim f(\theta_k), \quad \theta_c \sim p_0.$$

The model generating the labels  $c_{j,\ell}$ 's may be reformulated as

$$\lambda_0 \mid \alpha_0 \sim \text{Dir}(\alpha_0/C, \dots, \alpha_0/C),$$

$$G_j = \sum_{k=1}^C \lambda_j(k)\delta_k, \quad \lambda_j \mid \alpha, \lambda_0 \sim \text{Dir}(\alpha \lambda_0), \quad c_{j,\ell} \mid G_j \sim G_j.$$

Another representation is given by

$$G_0 = \sum_{k=1}^{C} \lambda_0(k) \delta_k, \quad \boldsymbol{\lambda}_0 \mid \alpha_0 \sim \operatorname{Dir}(\alpha_0/C, \dots, \alpha_0/C),$$

$$G_j = \sum_{\tau=1}^{\infty} \widetilde{\lambda}_j(\tau) \delta_{\psi_{j,\tau}}, \quad \widetilde{\boldsymbol{\lambda}}_j \sim \operatorname{SB}(\alpha), \quad \psi_{j,\tau} \sim G_0, \quad c_{j,\ell} \mid G_j \sim G_j.$$

A CRF arising from this generative model is as follows. Corresponding to the J groups, imagine J restaurants, each with infinitely many tables but finitely many dishes  $C = \{1, \ldots, C\}$  on their globally shared menu. The  $\ell^{th}$  customer belonging to the  $j^{th}$  group enters restaurant j, sits at a table  $\tau_{j,\ell}$ , and is served a dish  $c_{j,\ell}$ . While the restaurant assignments are predetermined by group memberships, the table assignment for the  $\ell^{th}$  customer in restaurant j is chosen as  $\tau_{j,\ell} \sim \widetilde{\lambda}_j$ , and each table  $\tau$  is assigned a dish  $\psi_{j,\tau} \sim \lambda_0$ . Customers sitting at the same table thus all eat the same dish. Multiple tables may, however, be served the same dish, allowing two customers enjoying the same dish to be seated at different tables. Given  $c_{j,\ell}$  and the corresponding table assignment  $\tau_{j,\ell}$ ,  $\psi_{j,\tau_{j,\ell}} = c_{j,\ell}$ . See Figure S.1.

Let  $n_{j,\tau}$  denote the number of customers in restaurant j at table  $\tau$ ,  $n_j(\psi)$  denote the number of customers in restaurant j eating the dish  $\psi$ , and  $n_j$  denote the total number of customers in restaurant j. Also, let  $n_{j,\tau}(\psi)$  denote the number of customers in restaurant j at table  $\tau$  eating dish  $\psi$ . Clearly,  $n_{j,\tau}(\psi) > 0$  only when dish  $\psi$  is served at an occupied table  $\tau$ . Finally, let  $m_j(\psi)$  be the number of tables in restaurant j serving dish  $\psi$ , and  $m_j$  be the total number of occupied tables in restaurant j.

Given a posterior sample of the dish assignments, we can obtain a draw from the posterior

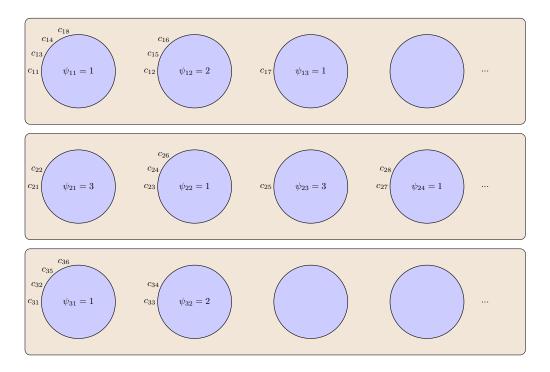


Figure S.1: The Chinese restaurant franchise.

of  $\lambda_j$  by noting that a-priori  $\lambda_j \sim \text{Dir}\{\alpha\lambda_0(1), \dots, \alpha\lambda_0(C)\}$  and that  $c_{j,\ell}$  for each  $j,\ell$  is a draw from  $\lambda_j$ . The number of different  $c_{j,\ell}$ 's that are associated with a specific dish c thus equals the total number of customers in the restaurant j eating the dish c, that is,  $n_j(c)$ . Using Dirichlet-Multinomial conjugacy, we then have

$$(\lambda_j \mid \alpha, \lambda_0, \mathbf{n}, \zeta) \sim \text{Dir}\{\alpha \lambda_0(1) + n_j(1), \dots, \alpha \lambda_0(C) + n_j(C)\}.$$

Likewise, given a sample  $(\tau, \psi)$  of the table and the dish assignments, we can obtain a draw from the posterior of  $\lambda_0$  by noting that a-priori  $\lambda_0 \sim \text{Dir}(\alpha_0/C, \dots, \alpha_0/C)$  and that  $\psi_{j,\tau}$  for each  $\tau$  is a draw from  $\lambda_0$ . The number of different  $\psi_{j,\tau}$ 's that are associated with a specific dish  $\psi$  is precisely the number of tables in the restaurant j that served the dish  $\psi$ , that is,  $m_j(\psi)$ . The total number of tables serving dish  $\psi$  across all restaurants is therefore  $m_0(\psi) = \sum_j m_j(\psi)$ . Using Dirichlet-Multinomial conjugacy, we then have

$$(\lambda_0 \mid \mathbf{m}, \boldsymbol{\zeta}) \sim \text{Dir}\{\alpha_0/C + m_0(1), \dots, \alpha_0/C + m_0(C)\}.$$

The table assignments  $\tau$  are also latent. To sample  $\tau$  from the posterior, we first

marginalize out their prior  $\widetilde{\lambda}_j \sim SB(\alpha)$  to obtain

$$(\tau_{j,\ell} \mid \alpha, \boldsymbol{\tau}_j^{-\ell}) \sim \sum_{\tau \in \mathcal{S}_{j,\tau}^{-\ell}} \frac{n_{j,\tau}^{-\ell}}{n_j - 1 + \alpha} \delta_{\tau} + \frac{\alpha}{n_j - 1 + \alpha} \delta_{\tau_{new}},$$

where  $n_{j,\tau}^{-\ell}$  denotes the number of customers sitting at table  $\tau$  in restaurant j excluding the  $\ell^{th}$  customer,  $\mathcal{S}_{j,\tau}^{-\ell}$  denotes the set of unique values in  $\boldsymbol{\tau}_j^{-\ell} = \{\tau_{j,s} : s = 1, \dots, n_j, s \neq \ell\}$  and  $\tau_{new}$  is a generic for any new value of  $\tau$  not in  $\mathcal{S}_{j,\tau}^{-\ell}$ . The distribution of the table assignment  $\tau_{j,\ell}$  given  $\boldsymbol{\tau}_j^{-\ell}$  and the dish assignments  $\boldsymbol{\psi}$  may then be obtained as

$$p(\tau_{j,\ell} = \tau \mid \psi_{j,\tau} = \psi, \alpha, \boldsymbol{\psi}_{j}^{-\ell}, \boldsymbol{\tau}_{j}^{-\ell}, \boldsymbol{\lambda}_{0}) \propto n_{j,\tau}^{-\ell} \delta_{\tau} \quad \text{if } \tau \in \mathcal{S}_{j,\tau}^{-\ell},$$

$$p(\tau_{j,\ell} = \tau_{new} \mid \psi_{j,\tau_{new}} = \psi, \alpha, \boldsymbol{\psi}_{j}^{-\ell}, \boldsymbol{\tau}_{j}^{-\ell}, \boldsymbol{\lambda}_{0}) \propto \alpha \lambda_{0}(\psi) \quad \text{if } \tau_{new} \notin \mathcal{S}_{j,\tau}^{-\ell},$$

where  $\psi_j^{-\ell} = \{\psi_{j,\tau_{j,s}} : s = 1, \dots, n_j, s \neq \ell\}$ . Since these assignments are restricted only to tables serving the dish  $\psi$ , the distribution reduces to

$$(\tau_{j,\ell} \mid \psi_{j,\tau_{j,\ell}} = \psi, \alpha, \boldsymbol{\psi}_j^{-\ell}, \boldsymbol{\tau}_j^{-\ell}, \boldsymbol{\lambda}_0) \sim \sum_{\tau \in \mathcal{S}_{j,\tau}^{-\ell}(\psi)} \frac{n_{j,\tau}^{-\ell}(\psi)}{n_j(\psi) - 1 + \alpha \lambda_0(\psi)} \delta_{\tau} + \frac{\alpha \lambda_0(\psi)}{n_j(\psi) - 1 + \alpha \lambda_0(\psi)} \delta_{\tau_{new}},$$

where  $S_{j,\tau}^{-\ell}(\psi)$  denotes the set of unique values in  $\boldsymbol{\tau}_j^{-\ell}(\psi) = \{\tau_{j,s} : s = 1, \dots, n_j, s \neq \ell, \psi_{j,\tau_{j,s}} = \psi\}$ ,  $n_{j,\tau}^{-\ell}(\psi)$  denotes the number of customers sitting at table  $\tau$  in restaurant j and enjoying the dish  $\psi$  excluding the  $\ell^{th}$  customer, and  $\tau_{new}$  is a generic for any new value of  $\tau$  not in  $S_{j,\tau}^{-\ell}(\psi)$ . This distribution can be identified with a marginalized conditional distribution of assignments of  $n_j(\psi)$  observations to different components in a SB $\{\alpha\lambda_0(\psi)\}$ . The full conditional for  $\lambda_0$  given  $(\psi, \tau)$  depends on the table assignments only via  $m_j(\psi)$  which can be obtained from the table assignments  $\tau_j$ .

Alternatively, for each of the  $n_j(\psi)$  customers in restaurant j enjoying the dish  $\psi$ , let  $m_{j,\ell}(\psi) = 0$  if the  $\ell^{th}$  customer sits at an already occupied table, and  $m_{j,\ell}(\psi) = 1$  if the  $\ell^{th}$  customer goes to a new table. Then,  $m_j(\psi) = \sum_{\ell=1}^{n_j(\psi)} m_{j,\ell}(\psi)$ . Using properties of a SB $\{\alpha\lambda_0(\psi)\}$  distribution, we then have

$$\{m_{j,\ell}(\psi) \mid \mathbf{m}_j^{\ell-1}(\psi), \alpha, \lambda_0\} \sim \frac{\ell-1}{\ell-1+\alpha\lambda_0(\psi)} \delta_0 + \frac{\alpha\lambda_0(\psi)}{\ell-1+\alpha\lambda_0(\psi)} \delta_1,$$

where  $\mathbf{m}_{j}^{\ell-1}(\psi) = \{m_{j,s}(\psi) : s = 1, \dots, \ell-1\}$ . We can then sample the  $m_{j,\ell}(\psi)$ 's from the posterior by sequentially sampling them as

$$\left[ \{ m_{j,\ell}(\psi) \}_{\ell=1}^{n_j(\psi)} \mid \alpha, \boldsymbol{\lambda}_0 \right] \sim \prod_{\ell=1}^{n_j(\psi)} \text{Bernoulli} \left\{ \frac{\alpha \lambda_0(\psi)}{\ell - 1 + \alpha \lambda_0(\psi)} \right\}.$$

## S.3.2 Higher Order CRF for CTF-HOHMM

While customers in the CRF of the HDP are pre-partitioned into restaurants based on their fixed group assignments, in our HOHMM setting the restaurant assignments are latent and hence are also sampled. Specifically, they are determined by the labels  $z_{j,t}$ 's - when  $(z_{1,t},\ldots,z_{q,t})=(h_1,\ldots,h_q)$ , the customer enters the  $(h_1,\ldots,h_q)^{th}$  restaurant. There are thus a total of  $\prod_{j=1}^q k_j$  restaurants.

We recall that the  $j^{th}$  lag  $c_{t-j}$  is important in predicting the dynamics of  $c_t$  only when  $k_j > 1$ . In the culinary analogy, the  $j^{th}$  lag is thus important if it has restaurants named (labeled) after it.

The total number of customers entering the  $(h_1, \ldots, h_q)^{th}$  restaurant is now  $n_{h_1, \ldots, h_q} = \sum_t 1\{z_{1,t} = h_1, \ldots, z_{q,t} = h_q\}$ . Among them, the number of customers eating the dish c is  $n_{h_1, \ldots, h_q}(c) = \sum_t 1\{z_{1,t} = h_1, \ldots, z_{q,t} = h_q, c_t = c\}$ . Using Dirichlet-Multinomial conjugacy, we then have

$$(\boldsymbol{\lambda}_{h_1,\ldots,h_q} \mid \boldsymbol{\zeta}) \sim \text{Dir}\{\alpha\lambda_0(1) + n_{h_1,\ldots,h_q}(1),\ldots,\alpha\lambda_0(C) + n_{h_1,\ldots,h_q}(C)\}.$$

We next define, for each  $\ell = 1, \ldots, n_{h_1, \ldots, h_q}(c)$ ,  $m_{\ell, h_1, \ldots, h_q}(c) = 0$  if the customer sits at an already occupied table and  $m_{\ell, h_1, \ldots, h_q}(c) = 1$  if the customer goes to a new table. Then, we can sample  $\{m_{\ell, h_1, \ldots, h_q}(c)\}_{\ell=1}^{n_{h_1, \ldots, h_q}(c)}$  from the posterior by sampling them sequentially from

$$\left\{m_{\ell,h_1,\dots,h_q}(c)\right\}_{\ell=1}^{n_{h_1,\dots,h_q}(c)}\mid\boldsymbol{\zeta}\sim\prod_{\ell=1}^{n_{h_1,\dots,h_q}(c)}\operatorname{Bernoulli}\left\{\tfrac{\alpha\lambda_0(c)}{\ell-1+\alpha\lambda_0(c)}\right\}.$$

Then,  $m_{h_1,\dots,h_q}(c) = \sum_{\ell=1}^{n_{h_1,\dots,h_q}(c)} m_{\ell,h_1,\dots,h_q}(c)$  gives the number of occupied tables serving the dish c in the  $(h_1,\dots,h_q)^{th}$  restaurant.

The table assignments in restaurants  $(h_1, \ldots, h_q)$  follow  $\lambda_0$ . Letting  $m_0(c) = \sum_{h_1, \ldots, h_q} m_{h_1, \ldots, h_q}(c)$  denote the total number of tables serving dish c across all such restaurants, we can update  $\lambda_0$  using Dirichlet-Multinomial conjugacy as

$$\{\lambda_0(1),\ldots,\lambda_0(C)\} \mid \zeta \sim \text{Dir}\{\alpha_0/C + m_0(1),\ldots,\alpha_0/C + m_0(C)\}.$$

# S.4 Sampling Prior Hyper-parameters

The full conditional for the hyper-parameter  $\alpha$  in the original CRF can be derived assuming a Ga(a,b) prior and adapting to West (1992). Following Antoniak (1974), integrating out  $\lambda_0$ , we have  $p(m_j \mid \alpha, n_j) = \alpha^{m_j} s^*(n_j, m_j) \Gamma(\alpha) / \Gamma(\alpha + n_j)$ , where  $s^*(n, v)$  are Stirling numbers of the first kind. Letting  $\mathbf{n} = \{n_j\}_{j=1}^J$ ,  $\mathbf{m} = \{m_j\}_{j=1}^J$  with  $v = \sum_{j=1}^J m_j$ , since the restaurants are conditionally independent, we have

$$p(\alpha \mid \mathbf{m}, \mathbf{n}, \boldsymbol{\zeta}) \propto p_0(\alpha \mid a, b) \ p(\mathbf{m} \mid \alpha, \mathbf{n}) \propto \exp(-\alpha b)(\alpha)^{a-1} \prod_{j=1}^{J} \left\{ (\alpha)^{m_j} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_j)} \right\}$$

$$\propto \exp(-\alpha b)(\alpha)^{a+v-1} \prod_{j=1}^{J} \left\{ \frac{(\alpha + n_j) \operatorname{Beta}(\alpha + 1, n_j)}{\alpha \Gamma(n_j)} \right\}$$

$$\propto \exp(-\alpha b)(\alpha)^{a+v-1} \prod_{j=1}^{J} \left\{ \left(1 + \frac{n_j}{\alpha}\right) \int r_j^{\alpha} (1 - r_j)^{n_j - 1} dr_j \right\}$$

$$\propto \exp(-\alpha b)(\alpha)^{a+v-1} \prod_{j=1}^{J} \left\{ \sum_{s_j=0}^{1} \left( \frac{n_j}{\alpha} \right)^{s_j} \int r_j^{\alpha} (1 - r_j)^{n_j - 1} dr_j \right\}.$$

Treating  $\mathbf{r} = \{r_j\}_{j=1}^J$ ,  $\mathbf{s} = \{s_j\}_{j=1}^J$  as auxiliary variables, we have

$$p(\alpha, \mathbf{r}, \mathbf{s} \mid \boldsymbol{\zeta}) \propto \exp(-\alpha b)(\alpha)^{a+v-1} \prod_{j} \left\{ \left(\frac{n_{j}}{\alpha}\right)^{s_{j}} r_{j}^{\alpha} (1 - r_{j})^{n_{j}-1} \right\}.$$

The full conditionals for  $\alpha$ ,  $r_j$  and  $s_j$  are then obtained in closed forms as

$$(\alpha \mid \boldsymbol{\zeta}) \sim \operatorname{Ga}(a + v - s, b - \log r), \quad (r_j \mid \boldsymbol{\zeta}) \sim \operatorname{Beta}(\alpha + 1, n_j), \quad (s_j \mid \boldsymbol{\zeta}) \sim \operatorname{Bernoulli}\left(\frac{n_j}{n_j + \alpha}\right),$$

where  $\log r = \sum_{j=1}^{J} \log r_j$ , and  $s = \sum_{j=1}^{J} s_j$ .

To sample the hyper-parameter  $\alpha$  in the HOHMM setting, we mimic the derivations in the CRF and introduce auxiliary variables  $r_{h_1,\dots,h_q}$  and  $s_{h_1,\dots,h_q}$  for each  $h_1,\dots,h_q$ . Let  $\mathbf{n}_0 = \{n_{h_1,\dots,h_q}\}$ ;  $\mathbf{m}_0, \mathbf{r}_0, \mathbf{s}_0$  are similarly defined. It can then follows that

$$\alpha \mid \boldsymbol{\zeta} \sim \operatorname{Ga}(a_0 + m_0 - s_0, b_0 - \log r_0),$$
 $r_{h_1,\dots,h_q} \mid \boldsymbol{\zeta} \sim \operatorname{Beta}(\alpha + 1, n_{h_1,\dots,h_q}),$ 
 $s_{h_1,\dots,h_q} \mid \boldsymbol{\zeta} \sim \operatorname{Bernoulli}\left(\frac{n_{h_1,\dots,h_q}}{n_{h_1,\dots,h_q} + \alpha}\right),$ 

where  $m_0 = \sum_{y_t} \sum_{y_{t-1}} \sum_{h_1,\dots,h_q} m_{h_1,\dots,h_q}(c_t)$ , log  $r_0 = \sum_{h_1,\dots,h_q} \log r_{h_1,\dots,h_q}$ , and  $s_0 = \sum_{h_1,\dots,h_q} s_{h_1,\dots,h_q}$ .

Additionally, with an exponential prior  $\varphi_0 \exp(-\varphi_0 \varphi)$  on  $\varphi$ , its full conditional is

$$\varphi \mid \zeta \sim \exp\{-(\varphi_0 + \sum_j jk_j)\varphi\}.$$

In simulation experiments and real data applications, we set the prior hyper-parameters at  $a_0 = b_0 = 1$  and  $\varphi_0 = 2$ . Our experiences with numerical experiments suggest the results to be highly robust to these choices.

## References

- Alexandrovich, G., Holzmann, H., and Leister, A. (2016). Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 103, 423-434.
- Allman, E., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37, 3099-3132.
- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *Annals of Statistics*, 2, 1152-1174.
- Barron, A. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical report.
- Doeblin, W. (1938). Exposé de la Théorie des Chaínes simples constants de Markoff á un nombre fini d'États. Revue Math. de l'Union Interbalkanique, 2, 77-105.
- Gassiat, E. and Rousseau, J. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli*, 20, 2039-2075.
- Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. Springer Verlag, Berlin.
- Lindvall, T. (1992). Lectures on the Coupling Method. John Wiley & Sons, New York. Reprint: Dover paperback edition, 2002.
- Rio, E. (2000). Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. Comptes Rendus de l'Académie des Sciences-Series I-Mathematics, 330, 905-908.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101, 1566-1581.
- Thorisson, H. (2000). Coupling, Stationarity and Regeneration. Springer, New York.
- Vernet, E. (2015b). Posterior consistency for nonparametric hidden Markov models with finite state space. *Electronic Journal of Statistics*, 9, 717-752.
- West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. Institute of Statistics and Decision Sciences, Duke University, Durham, USA, Technical report.